South Dakota State University

Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

1974

Automated Recognition of Human Speech : A System Survey, a Problem Analysis, and a Frequency Shift Experiment

David Robin Scott

Follow this and additional works at: https://openprairie.sdstate.edu/etd

Recommended Citation

Scott, David Robin, "Automated Recognition of Human Speech : A System Survey, a Problem Analysis, and a Frequency Shift Experiment" (1974). *Electronic Theses and Dissertations*. 4760.
https://openprairie.sdstate.edu/etd/4760

This Thesis - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

AUTOMATED RECOGNITION OF HUMAN SPEECH:
A SYSTEM SURVEY, A PROBLEM ANALYSIS,
AND A FREQUENCY SHIFT EXPERIMENT

BY

DAVID ROBIN SCOTT

A thesis submitted
in partial fulfillment of the requirements for the
degree Master of Science, Major in
Electrical Engineering,
South Dakota State University

1974

SOUTH DAKOTA STATE UNIVERSITY LIBRARY

AUTOMATED RECOGNITION OF HUMAN SPEECH:

A SYSTEM SURVEY, A PROBLEM ANALYSIS,

AND A FREQUENCY SHIFT EXPERIMENT

This thesis is approved as a creditable and independent investigation by a candidate for the degree, Master of Science, and is acceptable for meeting the thesis requirements for this degree. Acceptance of this thesis does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Thesis Advisor Date

Head, Electrical Engineering Department Date

AUTOMATED RECOGNITION OF HUMAN SPEECH: A SYSTEM SURVEY, A PROBLEM ANALYSIS, AND A FREQUENCY SHIFT EXPERIMENT

Abstract

DAVID ROBIN SCOTT

Under the supervision of Professor G. D. Nelson

A survey of fifty-two automated speech recognition systems was performed. The main characteristics of each system were analyzed, compared, and reported. Each system was classified according to the basic approach taken to recognize human speech. The problems and contributions that every system within each approach possessed were also reported. Four major problems blocking automatic speech recognition were defined and existing solutions discussed. Finally, an experiment was performed to reduce one problem, the difference in voice pitch among speakers. The experiment revealed that a frequency shift was appropriate to normalize this difference among speakers. The nature of such a frequency shift was not clearly indicated.

ACKNOWLEDGEMENT

The author gratefully acknowledges the contribution made to this research effort by the National Science Foundation funds granted under its trainee program. He also wishes to thank the Electrical Engineering Department of South Dakota State University for additional financial support, office space, and computer programming funds. In addition, he would like to thank Dr. G. D. Nelson, Dr. R. S. Burke, Dr. R. J. Lacher, and the members of the electrical engineering staff at S. D. S. U. for their thoughtful discussions and helpful suggestions during the course of this research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 SIGNIFICANCE

The following investigation into the automatic recognition of speech by machine was undertaken so that the fundamental problem of recognizing normal speech could be better understood and more easily solved. The survey of automated speech recognition systems was made because various approaches taken in the past were considered valuable when attempting to improve the recognition of speech by machine. The major problems nested within this fundamental goal were analyzed because they were crucial to the understanding and emphasis of past speech recognition efforts. Finally, an experiment was performed so that actual experience could be obtained with speech, the quantity being recognized and analyzed. In addition, the experiment was conducted in order to help solve the problem of pitch differences among speakers' voices.

1.2 OBJECTIVES

This literature and experimental research was conducted in order to accomplish several objectives. First of all, the establishment of a historical background of speech recognition efforts was desired. This background would reveal the various basic approaches taken in the past to recognize speech. In addition, the disadvantages and

contributions of each system would be noted in order to recognize and analyze the problems inherent in recognizing speech. Secondly, an analysis of the major problems blocking automatic speech recognition was desired not only to investigate these problems, but also to evaluate the state of their solutions. Thirdly, an experiment was performed to reduce a specific part of one of those problems, the differences which exist in the voice pitch of speakers. Finally, a more intangible objective of this research was to provide a basis of experience and knowledge from which a speech recognition system could be designed and built.

1.3 ORGANIZATION

The body of this research is reported in Chapters two, three, four, and five. The second chapter provides a survey of speech recognition systems which not only characterized each system, but also reports the problems encountered and the solutions discovered by these investigators. Several other articles, surveying speech recognition systems, were written by Flanagan (1), Lindgren (2), and Pols (3). The speech recognition systems discussed in the literature review are grouped according to the features used to recognize speech. The five groups include template matching, spectral feature, distinctive feature, zero crossing, and formant. All the systems in these five approaches follow the block diagram of an automated speech recognition system shown in Figure 1.1. Some of the differences among each approach are displayed by the alternate blocks shown for each approach. These five approaches, however, are not distinctly different in that they

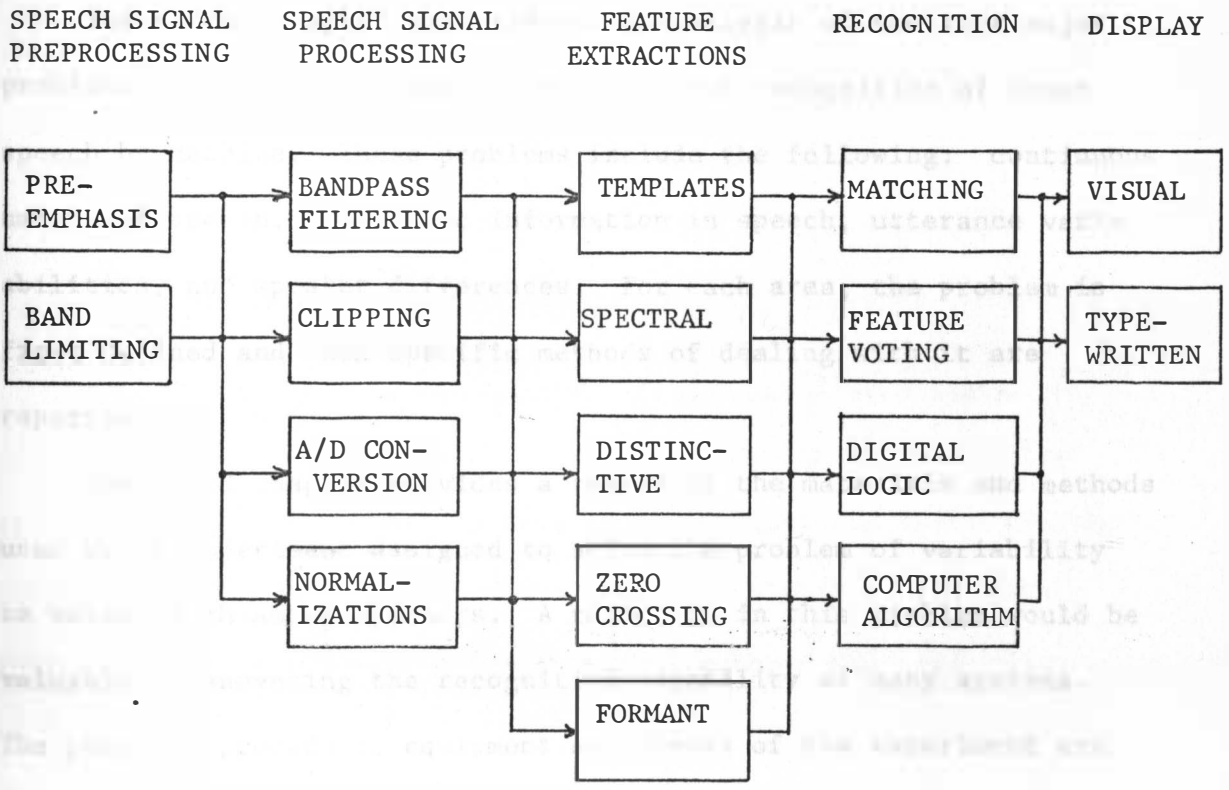

FIGURE 1.1 Block diagram of automated systems for recognizing speech.

tend to overlap when several techniques are combined in a recognition system.

The second chapter also reports an analysis of the four major problems which have prevented the successful recognition of human speech by machine. These problems include the following: continuous nature of speech, linguistic information in speech, utterance variabilities, and speaker differences. For each area, the problem is first defined and then specific methods of dealing with it are reported.

The third chapter provides a report of the materials and methods used by an experiment designed to solve the problem of variability in voice pitch among speakers. A reduction in this problem would be valuable in enhancing the recognition capability of many systems. The purpose, procedure, equipment and theory of the experiment are all discussed in this third chapter. The fourth chapter provides a report of the results of this experiment while the fifth chapter discusses these results. The final chapters are concerned with the conclusions and recommendations for applications and further research. Following the list of references, Appendix A is provided and should be examined because certain terms used in the body of this research are defined there. Appendix B details the results of several statistical tests made in connection with the experiment results in Chapter Four.

CHAPTER TWO

LITERATURE REVIEW

2.1 SPEECH RECOGNITION SYSTEM CLASSIFICATION

2.11 Template Matching

2.111 Analysis

The most popular approach in the past twenty-seven years to the automatic machine recognition of speech was that of matching templates of the frequency content of speech with the input speech. Each system used a series of stored library templates, and each utterance was compared with the templates until a best fit was obtained. Almost all of the templates used in this approach made use of the basic spectral information contained in the speech wave. A template usually consisted of a matrix of values of frequency content in one of several bands versus one of several time periods for each utterance. Naturally, each system required at least as many templates as there were vocabulary words. This approach, therefore, inherently involved large storage elements. In all of these systems, the templates were obtained for each speaker using the system, in order that recognition of his utterances was effective. In Table 2.1, the 17 template matching systems grouped here were comparatively displayed in chronological order from 1947 until 1973. The vocabulary, number of speakers, and recognition accuracy made up the basic characteristics chosen for display with each system. These

TABLE 2.1 Basic characteristics of template matching speech recognition systems in chronological order

System authors	Date year	Vocabulary	Speakers	Accuracy per cent
Kersta in (4)	1947	10 digits	1 speaker	99.5
Fry, Denes (14)	1953	14 phonemes	1 speaker	44
Olson, Belar (15)	1956	10 words	1 speaker	98
Dudley, Balashek (16)	1958	10 digits	1 speaker	100
Sebestyen (17)	1960	10 digits	10 males	99
Petrick, Willet (18)	1960	10 digits	30 speakers	99
Denes, Matthews (5)	1960	10 digits	6 males	94
Shearme, Leach (19)	1968	32 words	10 males	90
Purton (6)	1968	10 words	5 male, 1 female	93
Plomp, Mimpen in (3)	1968	discrete	50 speakers	82.5
Cannon (7)	1968	7 vowels	10 males	95
Washizawa in (3)	1968	10 digits	male and female	97
Clark (20)	1970	10 digits	2 males	98
Clapper (8)	1971	16 words	11 male, 2 female	94.2
Warren (9)	1971	10 digits	-*	100
Pols (3)	1971	20 words	20 males	95.5
Ichikawa, Nakano, Nakata (21)	1973	10 digits	1 speaker	100

*A dash indicates that the corresponding information was either unavailable or inappropriate.

systems employed a wide variety of vocabularies including the first ten spoken digits, the phonemes, the vowel sounds, or other discrete words. The vocabularies and the number of speakers for each system were kept small; therefore, all but two of the recognition scores surpassed 90 per cent accuracy. Each system was unique, however, and achieved this accuracy in different ways.

2.112 Characteristics of Template Matching Systems

2.1121 Frequency-Amplitude-Time Templates

The first and most popular approach to speech recognition by machine, template matching, has been modified and varied in many ways since Kersta first applied it to a speech recognition task in 1947. David (4) reported that Kersta simulated a recognition system for the first ten spoken digits. By using 200 hertz (Hz) bandpass filters, sampling every 67 milliseconds (ms), and quantizing each sample into two levels, a black and white mosaic was formed in the frequency domain for each digit. This template was matched with the incoming utterances of one speaker at a recognition accuracy of 99.5 per cent.

Denes and Mathews (5) used a 17-filter time-normalized template matching system to achieve 94 per cent correct recognition for six male speakers. Without time normalization, recognition accuracy dropped to 87 per cent. One female speaker brought the recognition accuracy down even further to 75 per cent. In another system, a 36 by 30 master template was cross-correlated with one of ten possible input utterances by Purton (6). The largest correlation coefficient

indicated recognition of the speech input as one of the vocabulary words. The master template was formed from one utterance from each of five male and one female speaker. The average recognition rate was 93 per cent, but recognition of a female's voice was less accurate. Pols (3) reported that Plomb and Mimpen used a five by eight matrix from 24 bandpass filters and eight samples per utterance. Fifty speakers scored a recognition accuracy of 82.5 per cent in real time.

Cannon (7) developed a recognition system that utilized speech information from one-half of the fundamental frequency period. The human ear was the model for this electronic analog ear, a 24-section filter network with outputs at four frequencies. Each sample was autocorrelated and compared to the stored templates. Seven vowels spoken by ten males were recognized with an accuracy of 95 per cent.

Pols (3) reported that Washizawa presented a system to recognize the first ten spoken digits using six bandpass filters. A logarithmic time scale was used with the template, and an accuracy of 97 per cent was attained for both male and female speakers. Clapper's (8) sixteen word system featured a low two to three minute time to establish the template for each speaker. The utterances of eleven male and two female speakers were recognized with an accuracy of 94.2 per cent. Eight bandpass filters and six nonlinear time segments made up the eight by six templates used in the recognition process. Warren (9) made a computer-simulated recognition study of the ten digits and showed a recognition accuracy ranging from 93 to 100 per cent. Each template was aided in the recognition process by the simplicity of

infinitely clipped speech, by feedback, and by sequential information.

A commercial system for the automatic recognition of speech was announced by Shields Products (10). Their template matching system utilized digital circuitry in the construction of their templates. A ten word system would cost nearly 700 dollars as advertised. In addition, several other systems have been announced by IBM (11) and Threshold Technology (12).

2.1121 Intermediate Phoneme Templates

In 1953 two new factors were exemplified in a system reported by Fry and Denes (13). First of all, recognition was attempted on the phoneme level using 14 templates for the 14 phonemes investigated for one speaker. A list of the common English phonemes is given in Table A.3 in Appendix A. Word recognition was then performed on the sequences of phonemes. Later, they (14) reported that recognition results were discouraging (60 per cent and 24 per cent for phonemes and words, respectively.). To improve these scores, digram frequencies were incorporated into the recognition system. This addition of linguistic information increased phoneme recognition to 72 per cent and word recognition to 44 per cent. Another speaker using the improved system was recognized with an accuracy of 45 per cent accuracy for the recognition of his phonemes.

Another intermediate phoneme recognition system developed by Olson and Belar (15) used eight bandpass filters and a frequency-amplitude-time template. Words spoken by one speaker with a ten-word vocabulary were recognized with an accuracy of 98 per cent

correct word recognition accuracy. A system developed by Dudley and Balashek (16) used intermediate phoneme recognition in recognizing the first ten spoken digits. Ten 300 HZ bandpass filters and ten time samples were used for each template. Amplitude normalization and sequential phoneme durations were added to produce near perfect word recognition for a male speaker. Other male speakers reached 90 per cent recognition accuracy with practice. Even though phoneme recognition results were not nearly as accurate, the proper word could usually be selected because of the limited vocabulary.

2.1123 Transformed Templates

In 1960, a digital computer was used by Sebestyen (17) to calculate a linear transform of the frequency-amplitude-time space to achieve better recognition accuracy. Time normalization was used with 18 bandpass filters to help the system recognize the ten digits spoken by ten male speakers. The template in this system consisted of the optimal linear transformation as computed from several series of utterances. There were no errors in one group of 400 utterances, and the total recognition error was less than one per cent. A similar system reported by Petrick and Willet (18) involved 30 speakers and a stored template representation of each of the ten digits. Again, recognition error was less than one per cent.

Twenty Dutch words were the object of a recognition system reported by Pols (3). The template in this system consisted of a three-dimensional trace. This trace was formed from the outputs of one-third octave bandpass filters using level normalization. The

system achieved 70 per cent correct recognition for eleven female speakers, 93.8 per cent correct recognition for twenty new speakers, and 95.5 per cent accurate recognition for twenty reference speakers.

2.1124 Orthogonal Templates

Shearme and Leach (19) formed a template of orthogonal representations for each of a 32 word vocabulary using 20 bandpass filters. Amplitude normalization was employed to achieve a recognition accuracy of 90 per cent for ten male speakers. Another computer-aided template matching system used sequence information and an orthogonal basis for recognition. In Clark's (20) system, six utterances were averaged together to produce a master template for each of the ten digits. For two speakers, the recognition rate averaged over 98 per cent correct.

2.1125 Comparison of Templates

Finally, Ichikawa, Nakano, and Nakata (21) reported a comparative study on template matching techniques. One speaker and the first ten Japanese digits were the basis for comparing templates based on the power spectrum, cepstral analysis, linear prediction, autocorrelation, and partial autocorrelation. Recognition results for the previous methods were 100 per cent accuracy with 25 power spectrum parameters, 100 per cent accuracy with six cepstrum parameters, 77 per cent correct with seven linearly predicted parameters, 92 per cent correct with four autocorrelation parameters, and 100 per cent accurate with six partial autocorrelation parameters. These five methods are defined briefly in Appendix A.

2.113 Problems with Template Matching Systems

The template matching approach to the automated recognition of speech had many inherent disadvantages built into each system. As implied earlier, large vocabularies were highly impractical due to the large number of templates needed. Therefore, whether hardware or a computer was used, the storage requirements were too great. Pols (3) in an attempt to reduce this problem, reduced 17 dimensions down to just three in recognizing his 20 word vocabulary with the aid of a computer having a memory of 8,000 words. Purton (6) also recognized this problem when storing the 36 by 30 patterns for his ten templates.

A second major problem, due to the nature of the template approach, was that of speaker acceptability. Valuable time must be taken to establish the templates for each individual using the system. In addition, no other speakers may effectively use the recognition system without re-establishing the template for their own voice. Clapper (8) estimated this training time for his system at a low two or three minutes, while other systems took days to adequately train. In addition to lost time, a mixed group of speakers was especially difficult to recognize because of the marked differences in pitch between male and female voices. Denes and Mathews (5) reported that errors increased from six per cent up to 25 per cent when female speakers were used with their system instead of male speakers. Pols (3) reported that errors for eleven women were 30 per cent, while errors for 20 men were 6.2 per cent. A final problem hidden in the phrase,

speaker acceptability, was the limitation on the number of speakers acceptable to the systems. The template approach became ineffective as more speakers were averaged into the template. One small evidence of this problem was exhibited by the mediocre results, 82.5 per cent recognition accuracy, obtained for fifty speakers by Plomp and Mimpen in Pols (3). On the other hand, a high recognition accuracy, 99 per cent, was obtained for thirty speakers by Petrick and Willet (18).

Another major problem accounted for a small amount of residual error in almost every system. Every utterance of a word by a speaker was different, and varied widely depending on the condition of the speaker and the speaking environment. Thus, a certain amount of misrecognition would occur when varying utterances were compared with the same templates. This problem was minimized in most cases by choosing a small yet divergent vocabulary. A forced decision rule, where all utterances were assumed to be in the vocabulary, helped solve this problem, but then created another problem, that is, improper recognition of utterances not of the vocabulary. Therefore, no provision could be made to indicate whether or not an utterance was contained within the vocabulary.

A final problem plagued just about every recognition system independent of which approach was used. This malady involved the time between the end of the utterance and the confirmed recognition by the machine of that utterance. Naturally, every designer was anxious to reduce this time to practically zero, that is, real time. Clark (20) reported delays of 16 to 30 seconds per word, while Plomp and Mimpen

292567 SOUTH DAKOTA STATE UNIVERSITY LIBRARY

in Pols (3) reported real time recognition. Most of the other systems reported delays from one-third of a second up to five seconds. Therefore, time was added to the system designer's considerations of recognition methods.

2.114 Contributions of Template Matching Systems

The seventeen template matching systems grouped together here have contributed to the solutions of those problems previously discussed and others. Fry and Denes (14) introduced the concept of intermediate phoneme recognition; so that eventually, continuous speech could be recognized on the subword level. Further discussion was devoted to the special problems of continuous speech recognition in section 2.21. Olson and Belar (15) refined phoneme recognition, and finally, Dudley and Balashek (16) produced near perfect discrete word recognition using phoneme recognition prior to word recognition. At least for discrete vocabularies, however, intermediate phoneme recognition was later discarded in favor of recognition based on the entire word. Fry and Denes (14) also introduced the use of digram frequencies to help in the recognition process of phonemes, much as the human uses linguistic information to recognize speech. Dudley and Balashek (16) also added phonetic duration information in order to correctly identify the vocabulary words from the less reliable phoneme sequences.

Several systems experimented with normalization of various parameters to reduce speaker and utterance variabilities. Denes and Mathews (5) reported that errors in their system decreased from

13 to 6 per cent with time normalization. Clapper (8) also found time and amplitude normalization were helpful in reducing these variabilities. Ichikawa, et al. (21) believed that a new nonuniform time-pattern matching system practically solved the problems dealing with utterance durations. Most designers considered amplitude normalization a necessity when comparing an input utterance with a stored template. In a somewhat different interpretation of normalization, the template was considered. Denes and Mathews (5) first showed that a template, composed of utterances of the same word by many speakers averaged together, reduced errors due to speaker variability. By averaging the utterances, individual variations were minimized to some extent, and the overall recognition rate was increased. Clark (20) used a master template that was made by averaging the utterances of one word by one speaker. This master template was found to be superior to an equal number of templates. Such a master template reduced the variability of a speaker's utterances by averaging them together. In spite of these successes with normalization techniques, further processing was needed for good recognition accuracy.

Another factor involved in the recognition process was that of timing information. It was shown that the vowel duration gave much information about the consonants immediately following the vowel (22). Naturally, the timing information was invaluable when recognizing a small number of widely differeing words. Dudley, et al. (16) used the duration information of phonemes to improve word recognition. Clark (20) included studies on word durations in his system, but

found formant information more helpful. Further discussion on the use of formants in speech recognition is included in section 2.15 as well as in Appendix A. Timing information, though, was still of value in recognizing speech.

A comparative study (21) of five pattern matching techniques indicated that partial autocorrelation coefficients were superior to any of the five parameter sets used. Next in order came cepstral analysis, power spectrum, autocorrelation, and linear prediction. In addition, the partial autocorrelation coefficients were also the easiest to extract from the speech wave when compared to the other four parameter sets. Not only can these methods be applied to the template matching approach to the speech recognition problem, but also to other general areas of approach to the same problem.

2.12 Spectral Feature Systems

2.121 Analysis

Spectral features formed the basis for a second popular approach to the automated recognition of speech. Each system used features extracted from the speech wave in order to identify the input utterance. Each feature was compared to threshold values in order to accomplish separation of input sounds and to recognize them. In addition to spectral features, other features and information were employed in order to increase recognition accuracy. Naturally, the biggest difficulty in using this approach was the discovery and the extraction of suitable features. The fourteen systems grouped here demonstrated good recognition accuracy for discrete utterances, while slightly lower accuracy was obtained for continuous utterances. In Table 2.2, the basic characteristics of these fourteen spectral feature systems are displayed with the systems in chronological order.

2.122 Characteristics of Spectral Feature Systems

2.1221 Discrete Word Systems

A second popular approach to the speech recognition problem utilized features based on the speech power spectrum. In 1959, David (23) reported a system designed by Shultz to recognize the first ten spoken digits. The utterances of 25 men and 25 women were recognized correctly 97 per cent of the time. In 1961, twenty-one speakers' utterances were analyzed by the 35 bandpass filters set up by Keith-Smith and Klem (24). Amplitude normalization, word frequencies, and a feature-selecting algorithm were used to achieve 94 per cent accuracy.

TABLE 2.2 Basic characteristics of spectral feature speech recognition systems in chronological order.

System author	Date year	Vocabulary	Speakers	Accuracy per cent
Shultz in (23)	1959	10 digits	25 male, 25 female	97
Keith Smith, Klem (24)	1961	10 vowels	21 speakers	94
Halle, Stevens (30)	1962	continuous	-*	-
Nelson, Herscher, Martin, Zadell, Falter (32)	1967	16 phonemes	6 males	94.8
Reddy (33)	1967	1-2 seconds of speech	1 speaker	81
Gilli, Meo (25)	1967	10 digits	10 speakers	90
Alter (31)	1968	continuous	-	-
Comer (26)	1968	10 digits	14 speakers	96.8
Lavington (34)	1969	10 digits	14 male, 5 female	97
Becker in (3)	1969	16 words	1 speaker	94.7
Velichko, Zagoruiko (27)	1969	168 words	1 speaker	95.2
Paul (28)	1970	50 words 30 words	1 speaker 3 speakers	99 90
Miller, Ross, Wine (29)	1970	10 digits	1 male	89
Reddy, Erman, Neely (35)	1973	16 words	-	-

*A dash indicates that the corresponding information was either unavailable or inappropriate.

Gilli and Meo (25) used 17 one-third octave bandpass filters between 110 Hz and 5600 Hz to distinguish among the first ten Italian digits. Combinational and sequential logic were used to process the spectral information obtained by the bandpass filters. New utterances of ten speakers were recognized with 90 per cent accuracy. Seventy per cent correct recognition was obtained for the utterances of additional speakers for which the machine was not adjusted.

A highly accurate system dubbed the "shoebox" (26), utilized waveform asymmetry to indicate voicing. Positive asymmetry exists in the waveform of voiced speech when the peaks above the time axis are greater than those below the axis. The unvoiced spectrum does not seem to have these differences in the peaks like the voiced spectrum does. Other parameters had to be adjusted for each of 14 speakers to obtain 96.8 per cent correct recognition for a limited 16-word vocabulary. Polo (3) reported that sixteen Danish words formed the vocabulary for Becker's system. One speaker attained a correct recognition rate of 94.7 per cent, by using spectral features extracted from two-octave bandpass filters. In addition, transitions within words were examined to aid recognition.

Velichko and Zagoruiko (27) reported a 168-word system using only five bandpass filters and sampling at 14 ms intervals. One speaker who trained the system scored a 95.2 per cent word recognition rate. Paul (28) used parameters based on the power spectrum of speech to achieve 99 per cent correct recognition for one speaker using a 50-word vocabulary. Recognition was reduced to 90 per cent, however, when

three speakers used a 30-word vocabulary with the system. Finally, a system for recognizing the first ten digits was developed by Miller, Ross, and Wine (29) in 1970. Twelve filters, amplitude normalization, and context information were used to achieve 89 per cent accuracy. Twelve features were extracted and the minimum Hamming distance was used to indicate recognition. This system was made adjustable to any male speaker and made compatible with a time-shared computer.

2.1222 Intermediate Phoneme Spectral Feature Systems

Halle and Stevens (30) in 1962 proposed a spectral feature recognition system for continuous speech. Intermediate phoneme recognition and a preliminary analysis were the basis for recognition. Rules were formulated for the synthesis of a likely sequence of phonemes. This sequence was then compared and updated with speech input information. A control component was proposed to save time by comparing phonemes in their most likely order. A proposal by Alter (31) for recognizing continuous speech focused on linguistic information sources. Alter hypothesized that this information could be used to bring preliminary acoustic recognition up to satisfactory levels of accuracy.

Another intermediate phoneme system employed weighted spectral features extracted from 19 bandpass filters. Besides linguistic information, Nelson, et al. (32) used two parameters of special value, the positive and negative slopes of the speech wave. Ten vowels and six consonants spoken by six males were recognized with an accuracy of 94.8 per cent. Reddy (33) experimented with a computer-simulated continuous speech recognition system. The seven step recognition

process involved amplitude normalization, segmentation, pitch extraction, spectral envelope analysis, feature extraction, segment group classification, and segment identification. Segmentation was accomplished using simple zero-crossing and intensity measures. Pitch synchronous analysis gave a good measure of the spectral envelope, up to the 100th harmonic of the pitch frequency. Other parameters extracted were duration, formant frequency and amplitude, zero frequency and amplitude, and noise concentration frequency and amplitude. For 30 short one to two second utterances, 81 per cent of the phonemes were correctly identified for one speaker.

Lavington (34) achieved 97 per cent correct recognition for 14 male and five female speakers. A vocabulary of the first ten digits was subdivided into 16 syllables for recognition purposes. Intermediate phoneme recognition was not only based on spectral parameters and pitch, but also on the zero crossing rate and the number of zero time derivatives of the speech wave. Finally, the last system grouped here under spectral features was one proposed in 1973 by Reddy, Erman, and Neely (35). This system attempted to utilize acoustics, syntactics, and semantics in three parallel recognition processors. Extensive use of linguistics was aided by various spectral parameters to recognize a 16-word continuous vocabulary. Recognition was achieved in four to seven times real time with this computer-based system.

2.123 Problems of Spectral Feature Systems

The use of spectral features to achieve machine recognition of speech had many of the disadvantages common to the other approaches

considered. As with the template matching approach, large vocabularies were difficult to recognize with this approach. As more and more words were added to the vocabulary, more and more parameters had to be extracted to recognize the added words. This extra complexity is then multiplied in the processing and recognition stages of the system. Alter (31) reported that a large memory was needed to store the syntax information associated with a large vocabulary.

Another problem inherent in this approach was speaker acceptability. Some progress was made in extracting speaker-independent features, but the task is still formidable. Some systems reduced this problem by using self-adjusting thresholds for their various recognition parameters. In addition, many of the systems were able to achieve high recognition rates for both male and female speakers. Furthermore, this type of approach did not seem to seriously limit the number of speakers able to use the system.

A third major problem not only plagued the spectral feature approach, but also the four other approaches as well. The variability in utterances of the same word by the same speaker caused a small amount of misrecognition. Naturally, the majority of the extracted features tended to vary along with the utterances, and the widest variations caused the misrecognition. Thus, research is being directed towards finding speech features which were unaffected by the inherent variabilities in these utterances.

Recognition time was once again a troublesome problem faced by these fourteen spectral feature systems. Gilli and Meo (25) reported

real-time recognition while others reported up to 40 times real-time for recognition. Spectral features had an advantage over template systems, however, in that many time consuming comparisons were avoided. Naturally, time was important from a practical point of view so that recognition could be accomplished at a rate commensurate with the speaker's rate of speech.

A final problem inherent in this feature extraction approach was that of selecting appropriate features. The criteria for ideal discriminating features were the following:

1. easily measured or isolated
2. present in every spoken utterance
3. present for every speaker
4. the same for the same sound as spoken by different speakers
5. the same for each utterance of the same sound by one speaker
6. unaffected by any change in a speaker's voice due to health, noise environment, or aging
7. markedly different for different sounds or groups of sounds

Naturally, several features were necessary and must complement each other in order to adequately separate all the speech sounds. Unfortunately, most of the combinations of features found thus far have not been able to satisfy these criteria. Moreover, it was widely accepted that features which will always identify the spoken utterance correctly have yet to be found.

2.124 Contributions of Spectral Feature Systems

Even though these fourteen spectral feature systems have problems,

they have contributed much to speech recognition research. Intermediate phoneme recognition was refined in several systems to produce excellent results. Amplitude normalization was nearly universally accepted and used in the majority of these systems.

The biggest contribution of this group of systems was in the area of feature selection. Keith-Smith and Klem (24) used an algorithm based on statistical decision theory to select the features used in their system. Nelson and Levy (36) provided a model that would select optimum features depending on time and cost factors. The features selected, however, did not necessarily provide 100 per cent classification. Nelson, et al. (32) reported that the positive and negative derivatives of the speech waveform were more reliable than formant data. Reddy, et al. (35) also observed that spectral features produced better recognition results with his system than did formant data. Reddy (33) earlier found that measures of duration and intensity were equally as important as spectral information in the recognition process. He also noted that the high frequency or noise characteristics were a significant clue towards recognition. Nelson, et al. (32) used weighting with the features in their system in order to achieve higher recognition rates. Comer (26) discovered that waveform asymmetry parameters were important to determine whether a sound was voiced or unvoiced. Lavington (34) found that zero crossings per ten milliseconds were related to the first formant frequency if above 280 Hz, and the number of zero time derivatives was related to the average of the second and third formants. The

formants were related to speech recognition in the section 2.15 concerning formant systems.

In the broad area of linguistics, much progress was made in implementing context, syntax, semantics, and other linguistic constraints. Halle and Stevens (30) proposed a scheme to compare phonemes in the most likely order according to the context and a preliminary analysis. Alter (31) proposed an extensive linguistic system to augment acoustic recognition efforts. Finally, Reddy, et al. (35) reported making use of three processors to recognize speech. The acoustic processor used not only acoustic features, but also phonological context and the vocabulary. The syntactic processor was made to induce words left or right by examining the grammar. Thirdly, the semantic processor used the meanings of the words and the meaning of the phrase to achieve recognition. These three processors were then used in parallel to recognize the speech input, becoming the first speech recognition system to use nontrivial linguistic constraints.

Recognition by synthesis was the last contribution noted in this group of speech recognition systems. Halle and Stevens (30) proposed such a scheme in 1962. Feedback and a flexible set of rules were used to generate speech patterns internally. These internal patterns were constantly compared with the input until a best fit was obtained. Reddy, et al. (35) employed nearly the same technique in each of their three parallel processors. They believed that every source of information and knowledge must be used at each stage of processing in order

to correct the inherent errors involved in each speech processing stage. Naturally, this scheme implies extensive use of feedback, feedforward, and cross-connections to correct the errors and resolve the ambiguities. Final results, however, were not available for either of these two innovative systems.

2.13 Distinctive Feature Systems

2.131 Analysis

Another significant group of speech recognition systems, those based on distinctive features, was launched by Jakobson, Fant, and Halle (2). Lindgren (37) outlined the twelve distinctive features corresponding to the twelve oppositions a human can physiologically produce. These twelve oppositions were defined in Table A.1 in Appendix A. Every language of the world was thought to be built on these universal oppositions of speech on the acoustic level. Table A.2 indicated the pattern English phonemes made in the nine distinctive features associated with the English language. In a broader sense, however, distinctive features could mean any features directly related to the phonetics of speech. The five systems grouped in Table 2.3 employed these features to recognize speech. Naturally, these systems were very similar to the spectral feature systems discussed earlier, and many of the same problems, therefore, are found in these two approaches. Each system, however, attacked these problems and the main task of speech recognition in a unique fashion.

2.132 Characteristics of Distinctive Feature Systems

2.1321 Systems Using Oppositions

The first group to attempt to utilize the distinctive features given by Jakobson, Fant, and Halle, was composed of Wiren and Stubbs (38). A set of eleven decision blocks was hypothesized to identify the English phonemes. Two of the blocks involved voicing, and two more involved the diffuse/compact opposition. The other seven blocks

TABLE 2.3 Basic characteristics of distinctive feature speech recognition systems in chronological order.

System authors	Date year	Vocabulary	Speakers	Accuracy per cent
Wiren, Stubbs (38)	1956	phonemes	21 speakers	94
Hemdal, Hughes (39)	1967	19 phonemes	1 speaker	92
Bobrow, Klatt (41)	1968	100 words	males	97
Itahashi, Makino, Kido (40)	1973	13 words	1 male	92.3
		53 words	1 male	92.3
Lea (42)	1973	continuous	-*	-

*A dash indicates that the corresponding information was either unavailable or inappropriate.

corresponded with the rest of the distinctive features associated with the English language. Twenty-one speakers obtained a correct recognition rate for phonemes of 94 per cent.

A system developed by Hemdal and Hughes (39) employed the same distinctive features and advantages of a computer. Nonsense syllables were composed of ten vowels and nine dipthongs, and a majority of the consonants. These nonsense syllables were identified with these four oppositions: grave/acute, compact/diffuse, tense/lax, and flat/plain. A computer calculated the physical correlates of these distinctive features from the outputs of 35 bandpass filters sampled at a 60 Hz rate. Normalization and contextual information were also used to increase recognition up to 92 per cent for one speaker.

Itahashi, Makino, and Kido (40) reported an interesting recognition system based on the distinctive features. Common discrete Japanese words were used to form the limited vocabulary. Segmentation of an utterance was performed by using some of the nine features based on only four frequency bands. With rules, segmentation accuracy reached 88.6 per cent. Intermediate phoneme recognition was accomplished by matching an input feature matrix with a set of stored matrices. The first word recognition scheme used a word-dictionary and phonological rules. For a 13 word vocabulary and one speaker, recognition accuracy for phonemes was 42 per cent, for syllables 59.5 per cent, and for words 92.3 per cent. For a 53 word vocabulary, the recognition rate dropped to 79.2 per cent. A duration dictionary was then added to increase recognition accuracy up to 86.8 per cent.

Separately, however, the duration dictionary outperformed both of the other schemes and achieved 92.3 per cent recognition for 53 words.

2.1322 Systems Using Related Distinctive Features

One of the most successful systems reported in the literature was described in 1968 by Bobrow and Klatt (41). Their set of 15 distinctive features was derived from the outputs of 19 bandpass filters. Another set of 29 abstract features were derived so that a comparison could be made between the two types of features. The recognition algorithm was based on simple voting of the features. The vocabulary could contain up to 100 discrete words spoken by a male speaker. Several utterances of each vocabulary word were used to adjust the recognition system. Time was removed from each utterance by recording only the changes in a feature from sample to sample. Some timing information was retained, however, by one of the features. Recognition accuracy was 97 per cent for a 54-word vocabulary for both sets of features devised.

Another proposed recognition system was outlined by Lea (42). He advocated the use of linguistic information, as well as prosodic features, that is, fundamental frequency, speech intensity, and durations. He pointed out that the fundamental frequency was closely correlated with intonation and stress contours of speech, while speech intensity and durations corresponded with the syntax of the utterance. The general recognition scheme suggested was that of recognition by synthesis as discussed in the last section, spectral features.

2.133 Problems of Distinctive Feature Systems

Several specific problems were associated with the distinctive feature approach to the problem of automatically recognizing speech. In addition, many of the problems discussed earlier are applicable to these systems as well. Bobrow and Klatt (41) echoed other researchers in that their features were not speaker independent, and that recognition degraded with an increase in the vocabulary or in the number of speakers. Itahashi, et al. (40) reported a drop of 22.3 per cent in recognition accuracy when nine male speakers used his system instead of just one. Therefore, variabilities in speakers and utterances accounted for many of the problems encountered by the investigators using this approach.

The major problem, however, was found in dealing with the feature extraction process. Itahashi, et al. (40) said, "Acoustic correlates of this feature system have not yet been fully described." The biggest problem was, therefore, the extraction of the distinctive features. A second problem occured when considering the distinctive features themselves, because they did not adhere to the criteria given earlier for the ideal discriminating feature. This problem was implied by Fant when the need was expressed for normalization of speech to reduce speaker differences in these features (39). Thus, speaker and utterance variability problems must be compensated for by other parts of the system in order to achieve good recognition results.

2.134 Contributions of Distinctive Feature Systems

These five investigations have made progress in finding the solutions to the many and varied problems dealing with automated speech recognition. A major contribution to some of these solutions was the use of distinctive features in intermediate phoneme recognition. These features were based on the phonetic realities present in speech, and the recognition process was modelled after the human. In addition, there was evidence that distinctive features were more concise and consistent than abstract features (41).

Many of these systems used other techniques to aid their basic recognition scheme. Hemdal and Hughes (39) used amplitude normalization and a computer to extract the features for their system. Bobrow and Klatt (41) employed time normalization by removing those samples which had remained unchanged. They also were able to work with a large vocabulary without adverse results. Finally, Itahashi, et al. (40) found that the use of durations in recognizing phonemes and words was also very important when identification is sought.

The last group of contributions made by these systems came in the area of linguistic information. Hemdal and Hughes (39) used simple contextual constraints to distinguish between dipthongs and some consonant effects. He avoided further linguistic constraints by using a nonsense syllable vocabulary. Bobrow and Klatt (41) also used limited linguistic techniques by instructing their recognition algorithm to use redundancy, to ignore inconsistencies, and to correct errors. Itahashi, et al. (40) pointed out that the restrictions on

vocabulary size were nearly equivalent to the use of context in a recognition system. He added certain phonological rules to achieve better segmentation and higher recognition accuracies. Most significantly, however, Lea (42) proposed an approach totally immersed in linguistic information. He advocated the use of syntactic and semantic information prior to phoneme recognition as well as afterwards. Intonation, stressing, and phonological rules would all be intertwined to provide the best possible recognition performance. Thus, once again ideas for the use of linguistic information by a machine permeated the distinctive feature approach to speech recognition.

2.14 Zero Crossing Speech Recognition Systems

2.141 Analysis

The use of zero crossing information of the speech wave comprised the fourth general category of speech recognition techniques. The remarkable discovery, demonstrated by Licklider and Pollack (43), that infinitely clipped speech was still intelligible to the human ear, suggested the use of zero crossings in recognizing speech. The question remaining was whether there was enough information left after clipping for a machine to recognize speech.

The nature of zero crossing information in the time domain in relation to spectral information in the frequency domain suggested the utility of this approach in solving the speech recognition problem. There were at least three types of zero crossing measurements used in automated speech recognition. The first yielded a good measure of the first two formant frequencies. Ewing and Taylor (44) claimed that the first formant was approximated by the average zero crossing rate of the speech wave, while the second formant was nearly equivalent to the average zero crossing rate of the differentiated speech wave. The second type of zero crossing information was a measurement of all the times or distances between adjacent zero crossings, and was usually used in the analysis of vowels (45). Finally, some vowels and consonants were analyzed using the number of zero crossings exceeding a set duration threshold (45). These three types of measurements were used in the systems comparatively displayed in chronological order in Table 2.4. Next, each system's characteristics and capabilities will be described in greater detail.

TABLE 2.4 Basic characteristics of zero crossing speech recognition systems in chronological order.

System author	Date year	Vocabulary	Speakers	Accuracy per cent
Bezdel, Chandler (47)	1965	5 vowels	20 male, 20 female	94
Scarr (50)	1968	-*	-	-
Ewing, Taylor (44)	1969	10 digits	5 males	-
Bezdel, Bridle (48)	1969	15 words	30 males	96
Ito, Donaldson (46)	1971	9 phonemes	1 male	-
Biancomano (51)	1972	-	-	-
DeMori (49)	1973	10 digits	4 males	98

*A dash indicates that the corresponding information was either unavailable or inappropriate.

2.142 Characteristics of Zero Crossing Systems

2.1421 Formant-Related Systems

A system developed by Ewing and Taylor (44) attempted to recognize the first ten digits as spoken by five male speakers. Clipping of the speech wave at 12 dB was introduced in the first step of processing to increase the relative amplitude of consonants over vowels. Two zero crossing rates approximating the first two formant frequencies were then used as a basis for recognition. Cross-correlation of the formant pattern was rejected in favor of the minimum Euclidean distance between the input pattern and the stored pattern in order to achieve final recognition. The first formant was band-limited betweem 300 Hz and 1000 Hz, and the second formant was band-limited between 800 Hz and 4000 Hz in order to increase the clarity of each formant pattern making recognition easier. Recognition accuracies were excellent for each individual when he also made the stored patterns. Intervoice results, however, were not acceptable.

Ito and Donaldson (46) reported the results of their work with zero crossing rates and automated speech recognition in 1971. Certain vowels, fricatives, and stop consonants were the object of their study. A detailed evaluation of their formant-related zero crossing parameters was given in their report. They found that for one male speaker the vowels /i/, /u/, and /ʌ/ could be separated with the two parameters. The unvoiced fricatives /s/, /f/, and /ʃ/ could also be separated with the same parameters. More information was needed, however, to separate the unvoiced stops, /p/, /k/, and /t/.

2.1422 Systems Using Categories of Time or Distance with Zero Crossings

A system using the second type of zero crossing measurement mentioned earlier was described by Bezdel and Chandler (47). Each channel count represented all the zero crossing distances within the boundaries set for that channel. Three methods of comparing the six channel outputs to the stored reference set were used as a basis for recognition. The template, however, was established by a reference group of six women and four men. Simple cross-correlation matching produced 72 per cent correct recognition for the reference group and 67.5 per cent accuracy for the unknown group of 20 more women and 20 more men. Euclidean distance measurement between the input and the templates produced 88 per cent correct recognition for the reference group and 79 per cent for the unknown group. Weighting of the six channels was attempted using separately the standard deviation and the mean of each channel. The results for the reference group were 94 per cent and 84 per cent correct, respectively.

In 1969, Bezdel teamed with Bridle (48) in reporting another system utilizing zero crossing information. After pre-emphasis, the speech wave was bandpass filtered according to the formant regions. An ac bias at a frequency much higher than speech was also added to help eliminate noise. The durations between zero crossings in each band were measured and used as the basis for the recognition of 12 individual sound categories. The boundaries categorizing these durations were made flexible so that preceding sounds could preset the boundaries in the most likely place. Recognition of these sounds was

accomplished through the use of hardware where the input was compared to various threshold values. Word recognition was then accomplished through a computer algorithm that evaluated the sound sequences previously recognized. Thirty male speakers, for which the thresholds were adjusted, achieved recognition of 96 per cent for the 15 word vocabulary consisting of the first ten digits, "cancel", "nought", "space", "repeat", and "start". Twelve other males achieved 91 per cent correct recognition.

A recent speech recognition system using zero crossing techniques was reported by DeMori (49) in 1973. The speech wave was low-pass filtered to 1100 Hz and high-pass filtered from 500 Hz. The time between zero crossings was measured and filed into one of eleven categories varying between seven millisecond and 0.1 millisecond duration. The number of durations in each category per 20 ms were used as a basis for recognition. A value was then computed by weighting the categories and adding the weighted number of zero crossings together for each time segment. These values were graphed with respect to time in order to form a characteristic plot. These plots were then compared by a computer algorithm both on local and broad levels. Final recognition was performed by a set of computer algorithms corresponding to the vocabulary words. Four male speakers achieved 98 per cent correct recognition after a few hours of training. In addition, recognition was accomplished with a delay of only 20 ms.

2.1423 General Zero Crossing Systems

Scarr (50) in his system proposed using zero crossing information from several bandpass filters to obtain spectral information from speech. No definite results were given, but a comparison to bandpass filter techniques showed that zero crossings provided better recognition results. Biancomano (51) later reported a general method for constructing a limited speech recognition system. His approach was based on a small discrete vocabulary, one speaker, and basic electronic hardware. He found that the amplitude level, the slope information, and some zero crossing information were adequate to perform the desired recognition tasks.

2.143 Problems of Zero Crossing Systems

The first basic problem encountered when using zero crossing information was whether or not there was enough information there to recognize speech. Ito and Donaldson (46) expressed that for certain stop consonants, contextual information was needed in order to recognize them. When zero crossings were used to approximate the formant frequencies, the question turned to whether the formant frequencies were adequate to recognize speech. For instance, Ewing and Taylor (44) reported that his system could not adequately identify the first ten digits using zero crossing rates alone. Thus, information other than zero crossings had to be used to achieve high recognition rates.

The second basic problem involved the inherent variabilities among speakers. Bezdel and Chandler (47) were the only experimenters to include women speakers in the recognition system, and women only

accounted for 61 of the 135 errors made in 600 utterances. Other investigators avoided using female speakers in the design of their systems. Of course, because several of these systems also used a template matching approach, the same variability problems involved with the template approach were present in these systems.

2.144 Contributions of Zero Crossing Systems

Even though comparatively few problems plagued these zero crossing systems, there seemed to be many contributions of this particular approach to the goal of automated speech recognition. First of all, there were several inherent advantages when using zero crossing information. Because it was essentially digital in nature, it made processing much easier than with spectral information, and because it was located in the time domain, computer time-sharing techniques could be used. Finally, because the zero crossing information corresponded closely with infinitely clipped speech, amplitude normalization was not even needed in most cases.

One of the most important contributions was made because Scarr (50) showed that broadband filtering prior to zero crossing analysis helped extract the spectral information more closely than without the prefiltering. Bezdel and Bridle (48), Ewing and Taylor (44), and DeMori (49) all used some form of prefiltering as suggested by Scarr (50). Even earlier, Bezdel and Chandler (47) suggested the use of two filtered bands in order to achieve better recognition.

In relation to the more popular approaches using filters, zero crossing information had several more advantages. Scarr (50) compared

the two methods, and he was convinced that even though less information was available by using zero crossing techniques, the zero crossing information was more compact and usable than the spectral information. Ito and Donaldson (46) indicated that the zero crossing information was more independent of speaker characteristics than was spectral information. Finally, Bezdel and Chandler (47) noted that without a series of filters, automatic adaptation to a speaker's voice would be easier to implement.

Another attempt to reduce variations among speakers was made by DeMori (49). He pointed out that the shape of his characteristic plots had few variations when comparing words spoken by different speakers. In another effort to reduce the same problem, Ewing and Taylor (44) employed clipping of the speech wave at 12 dB. This amount of clipping increased the relative amplitude of spoken consonants over spoken vowels. Because many of an individual's voice characteristics were contained in the vowels, 12 dB of clipping would also reduce the speaker variability. Thus, the use of zero crossing information to recognize speech contributed to solving not only the speaker variability problem, but also the general problem of automated speech recognition.

2.15 Formant Systems

2.151 Analysis

The last group of speech recognition systems used the formant frequencies of the speech waveform as a basis for their recognition. Formant frequencies were the points in the frequency domain where there was a concentration of speech energy or a vocal tract resonance. These formants were not always easily definable and sometimes were not even present in the acoustic waveform of speech. Peterson and Barney (52) pioneered the evaluation of formant frequencies in studying ten vowels produced by 76 speakers. Table 2.5 outlines the average fundamental and first three formant frequencies found for these vowels. Separate average frequencies were given for the 33 men, the 28 women, and the 11 children. Another type of formant frequency, the single equivalent formant (SEF) frequency was developed by Focht (53). The SEF frequency was related to the vowel perceived when a single formant was used as the stimulus. As shown in Figure 2.1, Durst and Lefkowitz (54) showed that the SEF frequency took on the value of the first or second formant, depending upon which was dominant. When neither of the two formants dominated, however, the SEF frequency was approximately the average of the first two formant frequencies.

The formant frequencies have been very valuable to the automated recognition of speech sounds. In reality, the formants could be visualized as another set of distinctive features which were extracted from speech to recognize that speech. Davis, et al. (55) said, "It is well established that the formant frequencies, particularly of first

TABLE 2.5 Averages of fundamental and first three formant frequencies for the vowels spoken by 33 men, 28 women, and 11 children, in hertz. Used by permission (52).

		Vowels									
		i	I	ɛ	æ	ɑ	ɔ	U	u	ʌ	ɝ
Fundamental (F0)	M	136	135	130	127	124	129	137	141	130	133
	W	235	232	223	210	212	216	232	231	221	218
	Ch	272	269	260	251	256	263	276	274	261	261
First Formant (F1)	M	270	390	530	660	730	570	440	300	640	490
	W	310	430	610	860	850	590	470	370	760	500
	Ch	370	530	690	1010	1030	680	560	430	850	560
Second Formant (F2)	M	2290	1990	1840	1720	1090	840	1020	870	1190	1350
	W	2790	2480	2330	2050	1220	920	1160	950	1400	1640
	Ch	3200	2730	2610	2320	1370	1060	1410	1170	1590	1820
Third Formant (F3)	M	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
	W	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960
	Ch	3730	3600	3570	3320	3170	3180	3310	3260	3360	2160

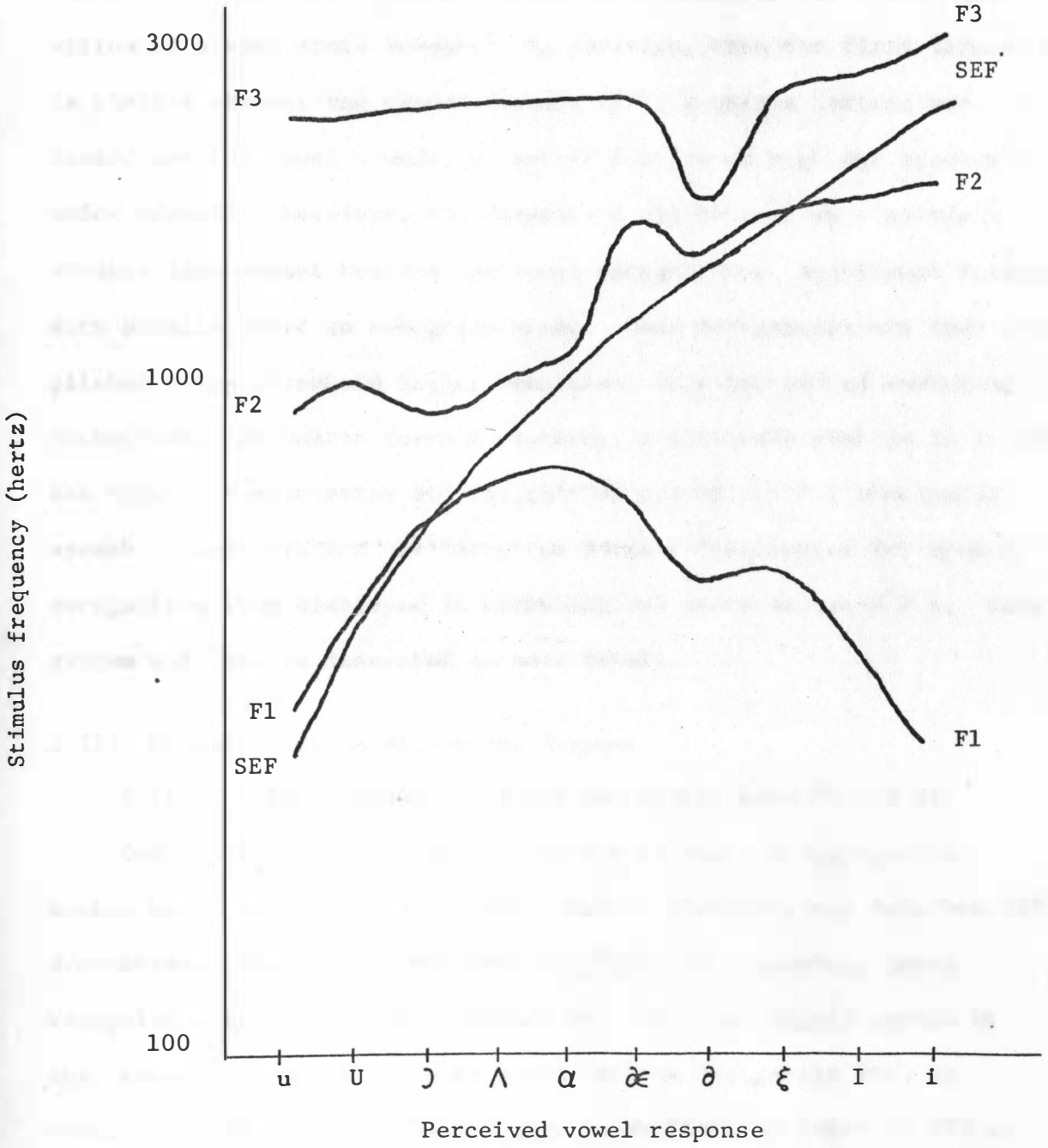

FIGURE 2.1 Positions of the three formant frequencies, F1, F2, and F3, in relation to the single equivalent formant frequency, SEF, for ten vowels. Used by permission, (54).

and second formants, serve as an important criterion in human recognition of steady state vowels." In addition, when the first formant (F1) is plotted against the second formant (F2), a unique pattern was formed for the vowel sounds, no matter how low or high the speaker's voice seemed. Therefore, the formants could be used as a partially speaker independent feature for vowel recognition. Additional features were usually added to recognize words. This recognition was then accomplished using threshold logic, templates, or other pattern-matching techniques. Automatic formant tracking, a difficult problem in itself, was used in segmentation and recognition procedures for continuous speech. These systems utilizing the formant frequencies for speech recognition were displayed in chronological order in Table 2.6. Each system will not be described in more detail.

2.152 Characteristics of Formant Systems

2.1521 Systems Recognizing the Ten Digits with F1 and F2

One of the first systems to use the formants in recognizing spoken words was unveiled in 1952. Davis, Biddulph, and Balashek (55) demonstrated that a machine, when adjusted for a speaker, could recognize with 97 per cent accuracy the first ten digits spoken by that speaker. Another speaker could only be recognized with an accuracy of 50 per cent for the same vocabulary. A delay of 350 ms was the time required to complete the recognition process. This system could equally qualify as a template matching, zero crossing, or formant-based system. The first and second formants were both approximated by the number of zero crossings of the filtered speech

TABLE 2.6 Basic characteristics of formant speech recognition systems in chronological order.

System authors	Date year	Vocabulary	Speakers	Accuracy per cent
Davis, Biddulph, and Balashek (55)	1952	10 digits	1 speaker	97
Forgie, Forgie (58)	1959	10 vowels	11 male, 10 female	93
Peterson (60)	1961	continuous	-*	-
Teacher, Kellet, and Focht (53)	1967	10 digits	10 males	90
Gerstman (59)	1968	10 vowels	76 speakers	97.5
Ohlendorf, Coates (56)	1968	10 digits	11 male, 14 female	91.8
Durst, Lefkowitz (54)	1970	10 digits	2 males	68
Von Keller (57)	1971	10 digits	4 males	93.9
Li, Hughes, Snow (61)	1973	continuous	4 males	96.3

*A dash indicates that the corresponding information was either unavailable or inappropriate.

wave below 900 Hz and above 900 Hz, respectively. A plot was then made of F1 vs. F2, and a five by six matrix of frequency-quantized blocks was superimposed on the F1 vs. F2 plot. Whenever a block was entered by the trace, an output corresponding to that block would be present until the trace had moved elsewhere. Thus, this matrix provided not only a measure of the F1 vs. F2 trace, but also of the behavior of the F1 vs. F2 trace with time. Moreover, this matrix contained the information that was matched with the ten templates stored by the system until a best fit was obtained.

Ohlendorf and Coates (56) used 22 one-third octave filters to approximate the formant frequencies. Intermediate phoneme recognition was used prior to the recognition of the first ten digits. Fourteen female and eleven male speakers scored 99.5 per cent correct recognition and 8.2 per cent misrecognition. Von Keller (57) reported a speech recognition system for the spoken digits in 1971. Automatic formant tracking of F1 and F2 was first used to segment an utterance and secondly to recognize it. Each of the six possible segments was represented by the vector (beginning F1, ending F1, beginning F2, ending F2, maximum F1). These vectors were compared with stored reference vectors to determine recognition. Segment identification was 72.1 per cent correct, while one male speaker's utterances were recognized correctly 94.2 per cent of the time. The system achieved 93.9 per cent accurate recognition for four other male speakers who used this system.

2.1522 Vowel Recognition by Formants

Forgie and Forgie (58) attempted to recognize ten vowels identical to those given in Table 5, except for a different surrounding context. The outputs of 35 bandpass filters were sampled at 180 Hz to determine the location of the first two formant frequencies between 115 Hz and 10 KHz. Vowel recognition was based on the location of an utterance's point in the F1 vs. F2 plane shown in Figure 2.2. Each block was labeled as to the vowels contained within it. A digital computer coupled additional information such as voicing, spectral derivatives, fundamental frequency (F0), and the third formant (F3). Initially, eleven men and ten women reached an average correct recognition level of 88 per cent but this was increased to 93 per cent by the addition of vowel duration information.

In 1968, Gerstman (59) studied the vowel sounds collected by Peterson and Barney (52). This system rescaled the first two formant frequencies from zero to 999, so that a computer algorithm could more easily recognize the ten vowels under consideration. Recognition accuracy was 97.5 per cent for 1,368 utterances.

2.1523 Continuous Speech Recognition by Formants

In 1961, Peterson (60) proposed a complex system to recognize continuous speech. He proposed that the first of seven steps to accomplish recognition would be the extraction of basic acoustical features from the speech waveform. He listed these features as F1, F2, F3, the first two anti-resonances, the average speech power, and the three prosodic parameters referred to in the section on distinctive

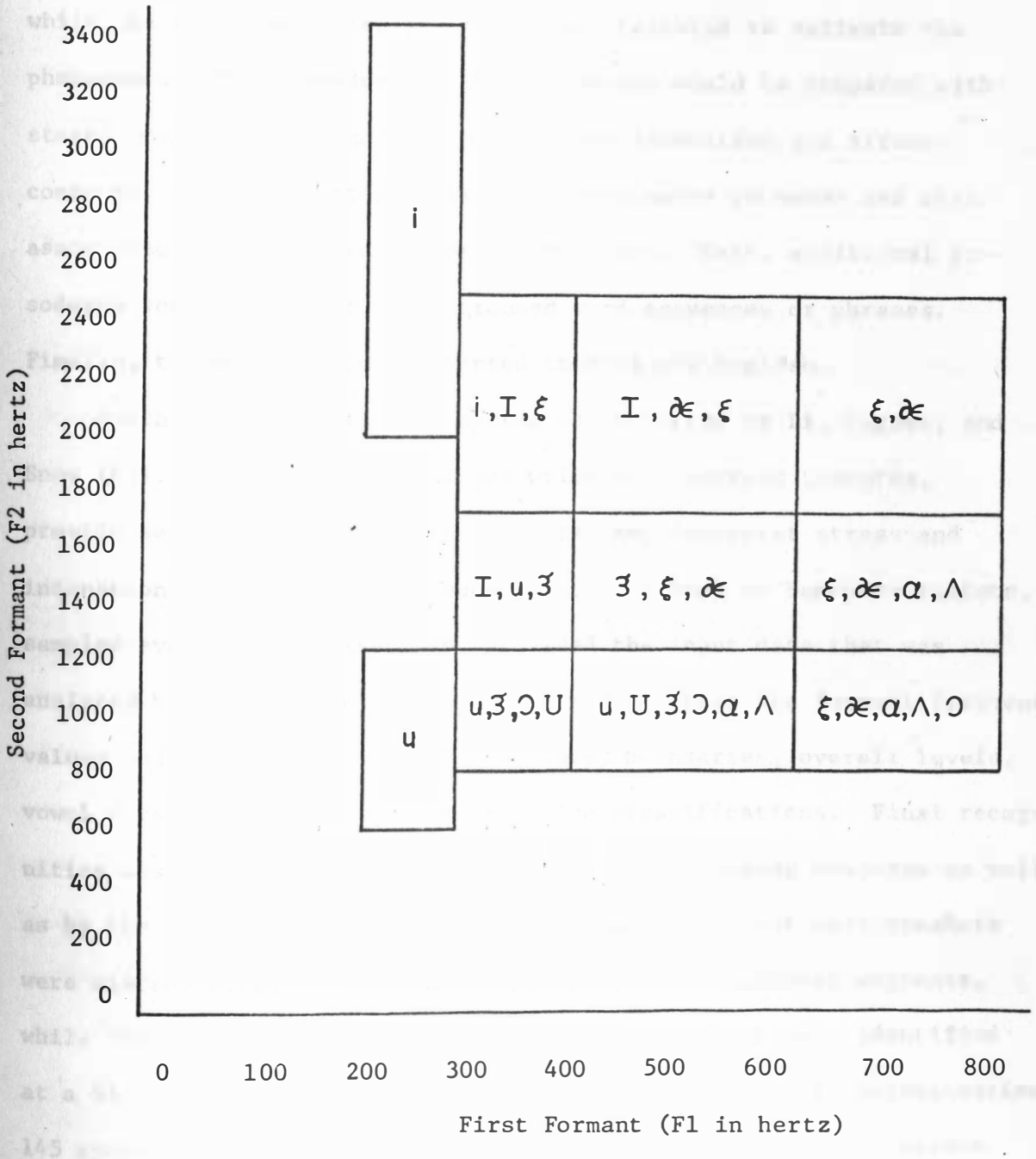

FIGURE 2.2 Location of vowels in the F1 versus F2 plane according to Forgie and Forgie (58). Used by permission.

features. The second step was the normalization of these features, while the third step used the normalized features to estimate the phonemes and the prosodemes. The prosodemes would be compared with stored prosodeme information to determine intonation and stress contours. The fifth step utilized the estimated phonemes and their associated probabilities to form word groups. Next, additional prosodemes would be used to form grouped word sequences or phrases. Finally, the words would be printed in ordinary English.

Another formant tracking system was reported by Li, Hughes, and Snow (61). A computer program was written to extract features, provide segmentation, classify segments, and interpret stress and intonation contours of continuous speech. A bank of bandpass filters, sampled every ten milliseconds, provided the input data that was analyzed by the computer. The computer calculated the formant frequency values and levels, fundamental frequency, boundaries, overall levels, vowel durations, and vowel and fricative classifications. Final recognition was done by the use of intonation and stressing measures as well as by the use of classified segment information. Four male speakers were misrecognized in less than 4 per cent of 160 uttered segments, while the stressed syllables uttered by six speakers were identified at a 95 per cent accuracy level. After a frequency-shift normalization, 145 spoken segments by one female speaker were correctly recognized 88 per cent of the time.

2.1524 Single Equivalent Formant Systems

The single equivalent formant was used to recognize the first ten

digits in a system reported by Teacher, Kellet, and Focht (53). Only three features, the SEF frequency, the SEF amplitude, and the state of voicing, were used to classify an utterance as one of the digits. The SEF frequency was approximated by one over twice the time between the glottal excitation and the first zero crossing thereafter. The frequency, amplitude, and voicing were quantized into 18, 8, and 4 levels, respectively. A stored set of sequential templates formed the basis for comparison with the input features. Ten male speakers achieved 90 per cent correct recognition, one per cent misrecognition, and nine per cent undecided. Results for 34 speakers were lower, due to reasons of dialect, decreased clarity, and unfamiliarity with the system.

The final formant system to recognize the first ten spoken digits was reported by Durst and Lefkowitz (54). The single equivalent formant frequency, the derivative of the SEF amplitude, and a state of voicing were extracted by a hardware system. These three features were examined at four selected sample points by threshold logic to make the final recognition decision. Two males achieved only 68 per cent recognition and 9.2 per cent misrecognition of their utterances. Four other speakers achieved even less accuracy with this system.

2.153 Problems of Formant Systems

Although many of the formant speech recognition systems were highly successful, many problems have yet to be solved before this approach can solve the problem of automated speech recognition. Many of the problems inherent in other approaches were evident when parts

of a system used that approach. For instance, those systems adopting the template matching approach encountered the problems related to that approach. In fact, every system struggled with the approach it took and had to justify that struggle by weighing the advantages and the disadvantages of that approach against the task chosen. The main struggle with the formant approach was tracking the formants.

The problem of automatic tracking of the formants was almost a subject in itself, due to the avid interest taken in it in the literature. Lavington (34) stated that not only was formant tracking difficult, but it was also very costly and time-consuming. In his analysis of the problem, the formants were unclear during 25 per cent of an utterance, which made formant tracking very difficult. In addition, Pols (3) remarked that present tracking techniques were all slower than real-time, with the possible exception of linear predictive methods. The discontinuities in Von Keller's (57) tracked formants were blamed on sudden changes in the formant frequency, interference by nasal resonances, and the combination of two formants into one. Finally, Lavington (34) noted that when words were spoken quietly, the harmonic content and formant structure was degraded even further. Thus, the problem of automatically tracking formants proved difficult, and attempts at its solution will be discussed in the next section.

Another problem inherent in the formant approach to speech recognition was the usefulness of the features, the formants. Although they were in one sense speaker independent, they did not meet the

criteria listed previously for ideal discriminating features. Pols (3) said that the formants were not adequate to separate vowels and that other speech sounds would be even more difficult to recognize. Moreover, it was reasonable to assume that more information and, therefore, more complexity was needed to recognize speech by the formant method.

A summary of the disadvantages of the formant approach to the recognition of speech by a machine was given in Table 2.7. The disadvantages of each of the four other approaches discussed previously were also briefly exhibited in the table. These disadvantages were the result of the problems encountered by each investigator in working with each method of approach.

2.154 Contributions of Formant Systems

There have been many techniques invented to automatically track the formant frequencies. Since the vocal cords and the shape of the vocal tract are responsible for producing the formant frequencies of speech, if that shape were known, then the formant frequencies could be computed. Schafer and Rabiner (62) outlined the following techniques for tracking the formant frequency locations: location of the peaks in the short-time amplitude spectrum, pitch synchronous analysis, low-order spectral-moment analysis, analysis by the synthesis of the speech waveform, and finally their own cepstral analysis. Atal and Hanauer (63) proposed a system based on linear prediction to track these formants and find their bandwidths. Li, et al. (61) reported that they used the first method, location of peaks in the frequency spectrum by a computer. One of their refinements included a procedure to retrace

TABLE 2.7 Disadvantages of each appraoch to the automated recognition of human speech.

Approach	Characteristic disadvantages
template matching	1. Large storage elements are usually needed. 2. It doesn't accept a large number of speakers. 3. It doesn't accept large vocabularies because of the storage and overlap created. 4. It doesn't accept widely varying utterances.
spectral feature	1. The ideal spectral features have not been discovered. 2. The larger the vocabulary, the more features are needed.
distinctive feature	1. Recognition accuracy degrades with an increase in the vocabulary. 2. Variabilities may be more pronounced in these features. 3. All the features have not been successfully extracted.
zero crossing	1. There is a loss of information due to the infinite clipping of the speech wave. 2. Zero crossing information is not directly related to the speech sounds.
formant	1. The presence of formants in unreliable. 2. The formants are difficult to track. 3. Formants do not provide enough information for recognition of all speech sounds.

and enhance the first three formant tracks. Von Keller (57) also used peak-picking techniques with the spectral-moment calculations confined to narrow frequency bands around the last formant value. Thus, many techniques have been found to help solve this formant-tracking problem.

In another area, two systems contributed their ideas to the problems encountered when both men and women were recognized. Forgie and Forgie (58) used a measure of voice pitch to normalize the difference between male and female speakers. Li, et al. (61) used a fixed frequency multiplier, 1.16 for women and 1.37 for children. Even with this normalization, however, recognition accuracy for women dropped 8.3 per cent.

Progress was also noted with the use of sequential information. As noted earlier, sequential recognition was employed in Davis, Biddulph, and Balashek's (55) system. Durst and Lefkowitz (54) reported a sample point selecting method to select the same points relative to the different ways in which the utterance was spoken. Finally, Forgie and Forgie (58) reported their use of vowel duration information to increase their recognition rate from 88 per cent to 93 per cent.

Contributions were also noted in the field of linguistic information sources. Peterson (60) advocated the use of extensive linguistic information primarily on the phoneme level of speech. Stored information of the prosodemes was also advocated to provide stressing and intonation information. Similarly, Li, et al. (61) produced a computer program to detect intonation and stressing for an utterance. Errors

on well-stressed segments were less than five per cent. The results of the detection of stressed segments were used in addition to the formants to recognize the speech input.

A summary of the advantages of the formant approach to the speech recognition problem was given in Table 2.8. The advantages for the four other approaches discussed were listed there, too. The advantages in the table resulted from the basic theory and processes behind each approach. In addition, the discoveries made by each system within an approach contributed to the advantages of that approach.

TABLE 2.8 Advantages of each approach to the automated recognition of human speech.

Approach	Characteristic advantages
template matching	1. It is very accurate for a limited system with a small vocabulary and small number of speakers. 2. Templates are simple in design. 3. It takes advantage of some timing information.
spectral feature	1. Features are matched more closely with the sounds of speech. 2. Features can be selected which are not affected by variabilities.
distinctive feature	1. Features are directly related to speech sound qualities. 2. Features may be more concise than other features.
zero crossing	1. Zero crossing information is digital in nature. 2. Zero crossing information is located in the time domain. 3. Amplitude normalization is not usually needed. 4. Automatic processes are easier with zero crossing measurements. 5. Zero crossing information may be more compact and usable. 6. Zero crossing information may be more speaker independent than other features.
formant	1. Formants have proven to be valuable in vowel recognition. 2. Formants are directly related to actual speech phenomena, vocal tract resonances.

2.2 SPEECH RECOGNITION PROBLEM ANALYSIS

Every automatic speech recognition system had at least four basic problems in addition to those problems inherent in the approach taken to recognize speech. These four basic problems were alluded to in the literature survey in chapter two and were briefly the following:

1. speech is composed of many connected sounds and words
2. speech recognition by the human uses much more information than just the acoustic waveform
3. each utterance by a speaker speaking the same sound is different
4. each utterance of the same intended sound by different speakers is different

Therefore, in order to recognize speech nearly as well as the human does, these four basic problems must be solved or at least reduced to a manageable form.

2.21 Continuous Speech

2.211 Problem Definition

The first problem blocking automated recognition of human speech was the continuity of human speech. Natural speech is not composed of discrete words or phonemes, but every utterance is a unique combination of sounds - complex tones, breathing, noise, and silence - all connected together. Speech can be fast or slow, erratic or steady, and slurred or distinct. Each sound, however, must first be spotted in the continuous acoustic waveform. Boundaries between sounds must be identified before the sound can be classified as one of the phonemes or other speech sounds.

Even if the boundaries were extracted, however, the sounds couldn't be easily classified. Reddy (33) said that each sound in an utterance was dependent on the sounds preceding and following it. For each phoneme listed in Table A.3, a set of allophones existed corresponding to the possible variations in the phoneme due to its different sound environments. Thus, there were many sounds which were found to constitute a spoken utterance. This variety in sounds made identification and recognition by machine very difficult.

Even if the sounds could be identified accurately, however, there is no consistent correspondence between those sounds and words. Moreover, recognition hinged on words, as no meaning was associated with just the individual sounds that made up the words. Therefore, in addition to identifying the sounds, a continuous speech recognition system must use those sounds to identify words. Furthermore, this identification was especially difficult because of the many combinations of sounds that made up the same word.

But even if the words were all identified correctly, one difficulty still remained. The words must be picked out of a continuous chain of sounds. Word boundaries were the chief obstacles in being able to pick out these words out of a string of phonemes. In addition, each phrase or sentence had intonation and stress contours which affected the understanding of the spoken utterance. Thus, when a word was spotted in a sequence of sounds, some measure of stress and intonation was needed to accurately recognize the full meaning of the spoken word. Naturally, these measures contributed even more complexity to the problem of continuous speech recognition.

2.212 Contributions to a Solution of the Continuous Speech Problem

The complexity of the problem tackled in recognizing continuous speech was evidenced by the use of discrete vocabularies in more than three out of every four systems surveyed. Furthermore, several of the continuous speech systems were only proposals instead of working systems. Reddy (33) acknowledged that the task was presently considered extremely hard. Two systems stood out as pioneers in attempting to solve this problem, however. Reddy (33) was able to achieve an 81 per cent correct recognition rate for individual phonemes in a short one to two second utterance. Segmentation on the phoneme level was performed prior to the recognition of each segment. Li, Hughes, and Snow (61) reported that their system segmented each utterance, and recognized the segments at a rate of 96.3 per cent. Intonation and stressing contours were used in the final recognition of words.

The first step in reducing problems created by continuous speech was choosing sound boundaries. The basic strategy behind this selection of boundaries was to find the points in speech where abrupt changes or silence occur. For instance, Reddy (33) based his phoneme segmentation on both intensity and zero crossing rates. Formant frequencies were also used in these segmentation procedures. Both Von Keller (57) and Li, et al. (61) depended on shifts in the formant frequencies to help segment their utterances. Itahashi, et al. (40) reported phoneme segmentation accuracy for his system reached 88.6 per cent. Unfortunately, not all boundaries were abrupt and easy to identify.

The second step to compensate for the continuous nature of speech involved the many sounds which a recognizer must be able to classify into the proper group. In the case of phonemes, all of a phoneme's allophones had to be recognized as that phoneme, even though their characteristics were slightly different. Peterson (60) stated that the goal of his system was an approximate phonemic transcription where each allophone was grouped under its phoneme. Otherwise, his system would have been far too complex. In addition, Reddy (33) stated that there wasn't as yet any one characterizing feature for each phoneme. Thus, classification of every sound was not perfected because the features were not found that would do so.

Thirdly, the problem of forming words from a series of sounds was revealed. The solution to this problem was to choose word boundaries and specify possible phoneme sequences for each word. Word segmentation was easier than phoneme segmentation because the word boundaries were selected from the phoneme boundaries. Once the phonemes were identified and grouped into words, then these words became the subject of recognition analysis. Several systems were able to identify a limited vocabulary this way. Dudley, et al. (16) achieved 98 per cent recognition for a vocabulary of the first ten spoken digits. Bezdel and Bridle (48) specified in a computer algorithm the acceptable sequences of classified sounds which corresponded to each of their 14-word vocabulary. A vocabulary in normal speech, however, is very large, and it is, therefore, difficult to specify all the phoneme sequences needed to make up each word in the vocabulary.

In addition, these sequences would eventually start overlapping among similar words.

Finally, the problems of punctuation, intonation, and stress contours were evident. To really extract all of the information from the acoustic waveform, it was important to find measures of how it was spoken as well as what was spoken. Li, et al. (61) reported the use of the fundamental frequency and duration information to detect stress and intonation contours. For highly stressed words, errors in selecting these words were less than five per cent. Bobrow and Klatt (41) used stressing as one of their features in the recognition process. In their case, the sum and difference of several selected filters measured the stressings of words. Generally, however, there were no widely accepted measures found for stressing and intonation, much less, good methods to display them with the output of a speech recognizer to enhance understanding.

2.22 Linguistic Information in Speech

2.221 Problem Definition

The second major problem in speech recognition by machine stemmed from the desire to emulate the human speech recognition system. The human has remarkable capabilities in deciphering a spoken message beyond a simple acoustic analysis of the speech by the ear. Peterson in his article (60) said that the acoustic waveform of speech didn't contain sufficient information to accurately interpret the speech because both the speaker and listener have learned so much about the speech process and the language used. Denes and Mathews (5) suggested

a more interesting question, whether or not there was enough information in the acoustic wave of speech for a machine to accurately recognize that speech. Whatever the answer, linguistic information does play an important role in human speech recognition and, therefore, can be used by machines as well.

The human uses linguistic information so efficiently and in so many different ways, therefore, it was thought to be desirable for a machine to utilize this same information. First of all, the human recognition system operates on the sub-word level. Vast information about allowable and most probable sound sequences are neatly tucked away inside the brain. When the ear hears speech, it automatically consults the brain for this information to have it ready for final recognition. In addition, the brain stores this information for many languages, dialects, and familiar individuals. Moreover, this information is constantly being updated and new information stored with every utterance recognized.

Secondly and simultaneously, stored information regarding syllable sequences and meanings are used to recognize words and phrases. The vocabulary stored as the allowable sequences of syllables is the main tool used to identify the input speech. Naturally, the brain has also done various other recognition tasks like word and syllable segmentation. The vocabulary does not limit the identification process, though, because unknown words can usually be recognized from the listener's experience with other related words in the vocabulary, known syllables within the word, or the surrounding context of the word.

Finally, linguistic information is used on an even higher level. Acoustic information at all levels aids recognition, but stressing and intonation contours are especially helpful here. It was thought that the meaning of a phrase or sentence was often detected without identifying each word separately (40). This was not only done by piecing words together, but also by the listener's knowledge of the speaker, what he has been talking about, and how he said an utterance grammatically and emotionally. The intonation and stress contours gave acoustic aid in determining the tone of voice, the emotional state, and the mental state. The listener also received information from gestures, facial expressions, and the movement of the lips when the speaker is visible. A good example of the use of linguistic information occurs when any listener hears an only slightly familiar language or dialect. In a short time it is usually possible to communicate quite well, even without visual information. The problem facing speech recognition research, however, was storing and using this linguistic information to recognize speech automatically by machine.

2.222 Contributions to a Solution of the Linguistic Information Problem

Many designers have tried to incorporate linguistic information into their recognition systems. In one sense, every system with a limited vocabulary, where every utterance was classified as one of that vocabulary, was using linguistic information. Probably the first investigators to directly utilize linguistic information, however, were Fry and Denes (14). Digram frequencies, or the probability of one phoneme following another, were included in the final recognition

decision. Peterson's (60) system advocated the use of phoneme sequence information too, as did several others.

Probably the most significant approach to using linguistic information was reported by Reddy, et al. (35). Their "hearsay" system consisted of three parallel processors all using linguistic information to recognize the input speech. The acoustic recognizer combined acoustic information, phoneme context, and vocabulary constraints. The syntactic recognizer specified possible words due to the immediate context of the unknown word. Finally, the semantic recognizer based its recognition on the meaning of the preceding utterance and the task involved. These three processors then worked together to agree on final recognition of an utterance.

While "hearsay" was being developed, other researchers proposed various ways of utilizing linguistic information. Lea (42) proposed the use of syntactic and semantic information at all levels of speech, including the phoneme level. Alter (31) advocated the use of context and redundant speech information to correct and enhance recognition. Peterson (60) proposed the use of prosodemes, a measure of rhythm, stress, and intonation, for aiding the recognition of word groups. These proposals have in common the fact that some method must be found to efficiently use this linguistic information. Furthermore, still other sources of information that the human uses to recognize speech are available for recognition purposes when it is learned how to use them automatically with a machine.

2.23 Variability of a Speaker's Utterance

2.231 Problem Definition

A third problem created by the nature of speech was that each utterance of the same sound by the same speaker was variable. It was presumed that every utterance was unique and that it was virtually impossible for a speaker to duplicate an utterance. Peterson and Barney (52) reported that these variations in utterances were not really distributed randomly, but possess sudden shifts, breaks, and other fluctuations. The least variability in repeated utterances was observed when a word or sound was spoken naturally and in the same context. Another study revealed that when trying to speak clearly and exactly, even greater variations in saying the same sound or word were apparent (41). Thus, even when attempting to say exactly the same sound, these variations in the characteristics of that sound were still evident and troublesome.

Variations due to repeated utterances of the same sound or word were usually even more pronounced than when a speaker was not trying to duplicate that sound or word. First of all, the duration of sounds and words was not only varied by different speaking rates, but also by the intonation, rhythm, and stressing applied by the speaker (41). Secondly, stressing also affected the pronunciation of each sound within a word. For instance, Bobrow and Klatt (41) reported that consonants may be incompleted, released weakly, or reduced in high frequency content when present in an unstressed syllable. The problem was that there was no definite pattern of these events observable for

a speaker's utterances. Finally, the environment that surrounded a sound or word affected the pronunciation of that sound. Unfortunately, this same environment did not have a consistent effect on a sound or a word. Thus, not only do variations in utterances exist, but these variations do not always vary either randomly or consistently.

2.232 Contributions to a Solution of the Problem of Variabilities

Several ideas have been suggested to alleviate this problem of variabilities in repeated utterances. The use of a small vocabulary enabled a machine to match an input utterance with one of the vocabulary with little error despite the variability in those utterances. The larger the vocabulary became, the chance increased that the varied utterances of each word would overlap with utterances of other words. Itahashi, et al. (40) compared a 13-word and 53-word vocabulary, and found that recognition rates were 92.3 per cent and 79.2 per cent, respectively. Bobrow and Klatt (41) also observed the degradation of recognition results with an increase in their system's vocabulary. As discussed in the section on contributions to template matching systems, master templates were useful in reducing problems created by repeated variable utterances. Clark (20) found that such a master template improved recognition results by averaging noise and individual utterances.

Another effort to reduce the effects of these variations was to avoid them in recognition processes. Features were desired which did not exhibit these variations but were adequate to recognize speech. This approach assumed that the variable information was not important

to the recognition of the spoken words. Presumably, any information that was needed and was also variable had to be normalized by the machine to reduce the variations to a usable level. The primary emphasis, however, was to find those features which did not vary from utterance to utterance. Nelson, et al. (32) reported that the derivative of the speech wave was a more invariant feature than the formant locations. They attempted to recognize speech by making many simple decisions, the majority of which would be unaffected by existing variations. Other investigators, it seemed, included this problem in their attempts to solve the next problem, variations among speakers.

2.24 Variability among Speakers' Voices

2.241 Problem Definition

The main problem in speech recognition according to Bezdel and Bridle (48) was the variability among speakers' utterances when each utterance was intended to be the same sound or word. Lavington (34) concurred that more research was needed to determine the variations in speech so that features and their tolerances could be specified more accurately. Among speakers, variations in duration, intonation, rhythm, stressing, and context were even more pronounced than the same type of variation in one speaker's utterances. In addition, different speakers possessed widely varying speech habits, especially dialect, vocabulary, pronunciation habits, and speaking rate (41). Moreover, it is impossible for two speakers to produce identical utterances because of the differences in each speaker's vocal tract and other physical characteristics (41).

A big difference among speakers' voices could be attributed to the pitch of different voices. It is obvious that most women have higher-pitched voices while most men have lower-pitched voices than average. In addition, the voice pitch changes with age, health, and use. It was evident from examination of Table 2.5 that men's voices were consistently lower in frequency than women's voices which were lower than children's voices. In fact, for every vowel the average measured fundamental and formant frequencies for those 76 speakers followed the above pattern. This marked difference in the frequency range of each speaker produced problems in recognizing speech by reliance on spectral data. Approximately three out of every four systems surveyed avoided this problem by using only male voices for recognition experiments. Furthermore, about one out of every five systems surveyed used only one speaker to avoid any interspeaker differences, including those variations in voice pitch. Other researchers, however, have accepted the challenge of this variability problem and attempted to minimize or solve it, as discussed next.

2.242 Contributions to a Solution of the Variabilities among Speakers

One interesting idea to reduce individual speaker characteristics was based on the fact that most of these characteristics were contained in the vowel sounds. Ewing and Taylor (44) decided to employ clipping of the speech waveform to 12 dB to reduce the vowel content relative to the consonant content of speech, thereby nearly eliminating speaker influence. Unfortunately, they had to report that intervoice recognition results were not satisfactory.

Another popular approach to this wide variance among speakers was to avoid the variabilities. Reddy (33) reported that his system used speaker-dependent features when they were needed to resolve recognition problems. Many other systems required that each speaker train the recognition system. Much emphasis was also placed on discovering features which did not change with a change in speakers or their voices. Ito and Donaldson (46) found that zero crossing rates were more speaker independent than spectral features. Durst and Lefkowitz (54) claimed that the two single equivalent formant parameters were almost speaker independent as well. In addition, the location of the first two formants served as a partially speaker-independent feature for vowel recognition.

Another scheme to compensate for these variations among speakers saying the same word was mentioned in the contributions to template matching systems. Shearme and Leach (19) averaged the utterances of a word, spoken by many speakers, together to form their master template. This averaging process helped reduce errors due to noise and due to the variations mentioned previously.

Probably the most emphasis, however, has been placed on the normalization of the differences among male and female speakers. Almost every surveyed system reporting separate results for men and women reported lower recognition scores for women than for men. One exception was noted when Bezdel and Chandler (47) reported that errors for women were slightly less than those for men. Still other systems included women in their system, but did not report separate results.

Shultz (23), Forgie and Forgie (58), Gerstman (59), and Ohlendorf and Coates (56) reported recognition results greater than 90 per cent accurate when recognizing both women and men speaking. Two notable schemes were devised to cope with the differences in voice pitch. Forgie and Forgie (58) used a crude measure of pitch based on a combination of filters between 115 Hz and 365 Hz. This pitch measure was used to normalize the difference among speakers' voice pitch. Li, et al. (61) used a fixed frequency multiplication of 1.16:1 for women compared to men. This transformation still left the women's recognition rate at 88 per cent, 6.3 per cent below that of the men tested. Thus, significant attempts have been made at reducing the variabilities among speakers, but none have been very successful.

The characteristics of the problem with speaker differences was grouped with the problems of the continuous nature of speech, linguistic information in speech, and utterance variabilities in Table 2.9. The distinctive characteristics of each problem were displayed to summarize the attributes of these four major problems.

TABLE 2.9 Characteristics of the major problems encountered in automated recognition of human speech.

Major problem	Characteristic problems involved
continuous nature of speech	1. segmentation of sounds 2. classification or identification of sounds 3. identification of words from a string of sounds 4. separation of words
linguistic information of speech	1. allowable and most probable sound sequence information available for use 2. semantic information available for use 3. syntax information available for use 4. prosodic information available for use 5. vocabulary information available for use
utterance variabilities	1. variations in the power spectrum 2. variations in rhythm and stressing 3. variations due to the context of the utterance
speaker differences	1. voice pitch differences, especially between male and female voices 2. variations in rhythm and stressing 3. dialectical differences 4. vocabulary differences 5. pronunciation differences

CHAPTER THREE

EXPERIMENTAL MATERIALS AND METHODS

The materials and methods employed in this experiment will be reported in this chapter. This experiment was undertaken to study one of the variabilities among speakers when each speaker was intending to say the same word, syllable, or sound. Specifically, the differences in the power spectrum of the intended utterances spoken by different speakers were examined. The basic hypothesis stated that the power spectrum of an intended sound spoken by two speakers was similar, although it was shifted in frequency. This shift corresponded to the difference in frequency between the voice pitch of each speaker. Therefore, the possibility of normalizing this difference between the voice pitch of two different speakers was investigated in order to enhance the automated recognition of speech.

A simple experiment was designed to investigate the differences in voice pitch of different speakers. A flow diagram of the speech processing utilized in the experiment is given in Figure 3.1 The following discussion explains each section of this flow diagram. Ten male and ten female English speakers were randomly chosen at South Dakota State University as a typical population that may use a speech recognition system. These speakers were instructed to read a list of all the letters of the English alphabet. The letters

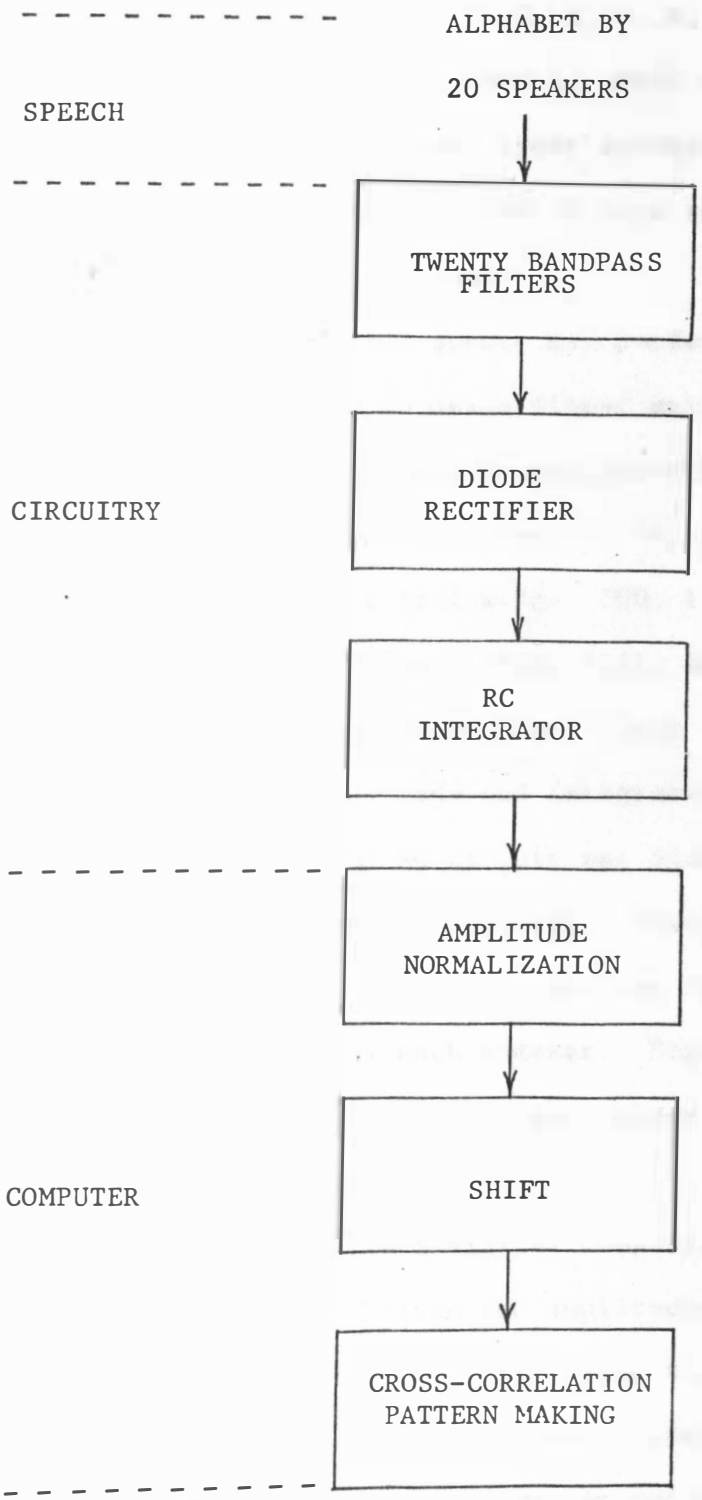

FIGURE 3.1 Experimental Flow Diagram

were read in the following order: V, P, J, D, W, Q, K, E, Z, T, N, H, B, U, O, I, C, X, R, L, F, Y, S, M, G, and A. Each speaker was instructed to speak naturally with a clear break between utterances. Their speech was recorded with an Akai X-1800 SD tape recorder in an acoustically untreated office environment.

The first step in processing this speech was performed by a model 3202 Krohn Hite filter. Each bandpass filter was adjusted so that the band between the cutoff frequencies was one-third of an octave, except for the last two filters. These 20 filters' lower cutoff frequencies in hertz were the following: 100, 133, 167, 200, 267, 333, 400, 533, 667, 800, 1067, 1333, 1600, 2133, 2667, 3200, 4267, 5333, 6400, and 8200. The output of each filter for each utterance was rectified by a biased diode and integrated over time by an RC circuit. The output of the RC circuit was displayed on an oscilloscope, and the final voltage was recorded. Thus, a set of 20 voltages, corresponding to each of the 20 bandpass filters, represented each utterance spoken by each speaker. Figure 3.2 shows the circuitry and equipment used to produce these 20 voltages representing each utterance.

Further processing was done with a digital computer. A computer program was written to linearly normalize the amplitudes of the 20-number sequences representing each utterance. After this normalization, the sequences representing each letter were cross-correlated with each other. In order to investigate the differences in voice pitch, the sequence representing the first voice was shifted with respect to the

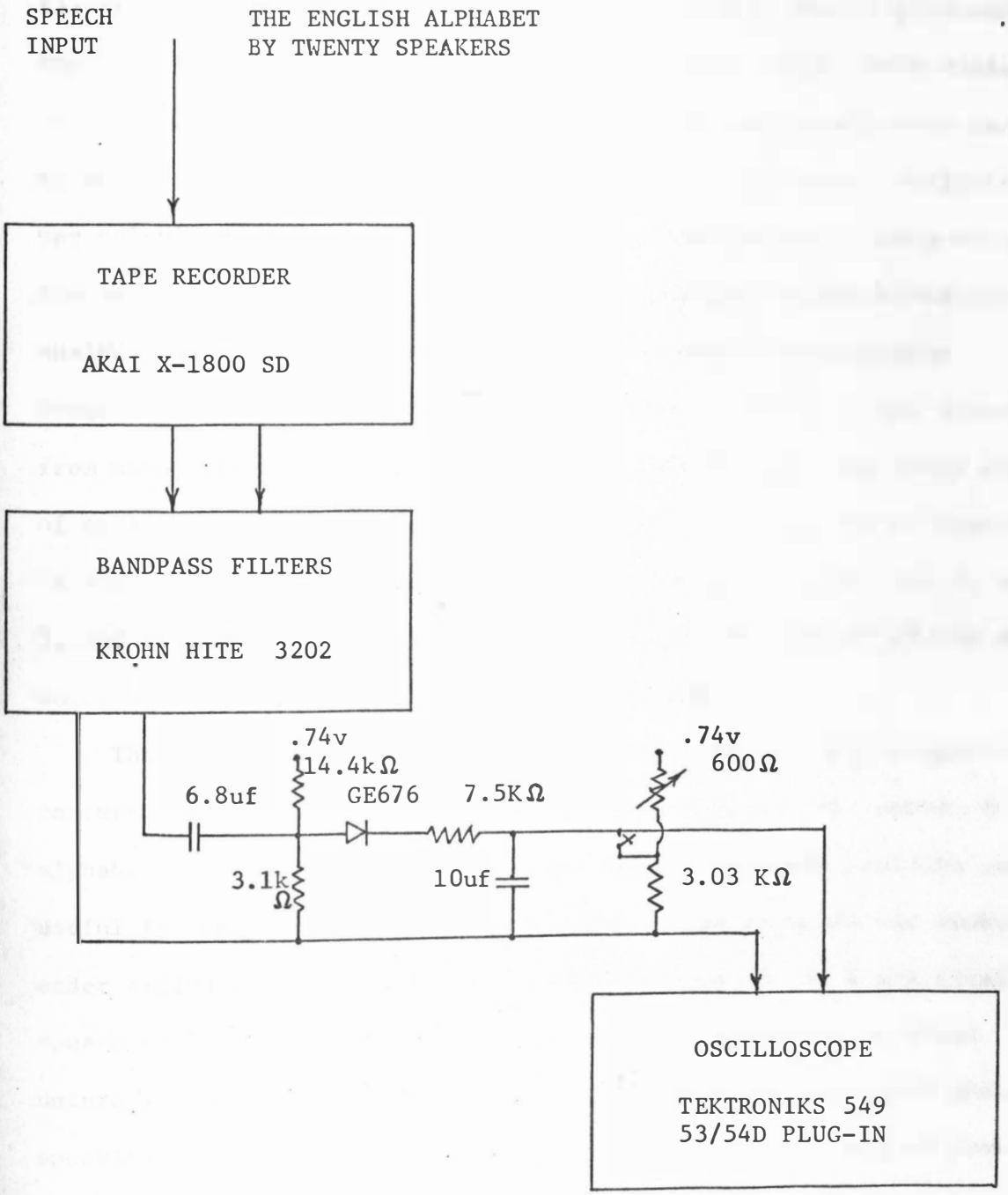

FIGURE 3.2 Equipment and circuitry used to process the input speech.

second sequence up to a distance of four filters positive and four filters negative. Zeros were inserted where these shifts produced empty spots on the ends of each sequence. These shifts were designed to simulate the raising or lowering of each voice's pitch with respect to another. Pearson's product-moment sample correlation coefficient was calculated as a measure of the linear correlation between every two sequences. Finally, these sample correlation coefficients were analyzed to investigate the hypothesis concerning interspeaker frequency differences. The peak correlation coefficient was chosen from among the nine possible correlation coefficients for every pair of speakers speaking each letter. This peak was recorded as occurring in one of the nine possible shift categories (-4, -3, -2, -1, 0, 1, 2, 3, and 4) representing the shift in the number of filters of the first voice while the second voice remained stationary.

The procedures outlined previously were adopted due to various concerns in addition to those previously explained. The spoken English alphabet was chosen as a medium-sized vocabulary which would be very useful for many speech recognition tasks. This alphabet was spoken under ordinary conditions such as could be expected in a practical speech recognition system. Each speaker was instructed to speak naturally so that the variability Bobrow and Klatt (41) noted when speakers enunciated carefully would be minimized. The filter bank was processing not only the input speech, but also the extraneous noise as well.

The choice of the one-third octave filters described previously was arbitrary and any configuration could have been chosen to investigate pitch differences among speakers. Several investigators used this type of bandpass filtering including Denes (64), Fry and Denes, (13), Gilli and Meo (25), Ohlendorf and Coates in Pols (3), and Pols (3). Pols (3) noted that one-third octave filters were used because they corresponded to certain physical measurements of speech. On the other hand, even though Denes (64) used one-third octave filters, he recommended using a set of bandpass filters with equal bandwidths below 100 Hz and bandwidths whose logarithms were equal above 1000 Hz. The bandpass filters used, however, were adjusted to include the normal range of speech frequencies, 100 Hz through 10 KHz. The output of each bandpass filter after rectification and integration represented the spectral density of the utterance in that filter bandwidth.

Processing of the outputs of the 20 filter channels continued in a digital computer. The amplitudes of the 20 outputs for each utterance were normalized to reduce variations due to the loudness and duration of each utterance. The normalized sequences for each spoken letter were cross-correlated with each other in order to provide a measure of the similarity of utterances. The sequences were shifted with respect to one another to simulate the raising or lowering of each voice's pitch with respect to another voice's pitch. These shifted sequences were also cross-correlated with each other to measure the similarity between utterances after shifting one of the voices with respect to the other. This type of shift was not linear

or uniform in design because of the increasing bandwidths of the filters. This type of shift, however, would be easy to implement automatically for a speech recognition system.

Two considerations led to the above approach to investigate speaker differences. It is well known that the pitch of a human voice poses no serious problem in the recognition of its speech by listeners. Therefore, the listener either normalizes speech according to the pitch, or else he used features in his recognition process that do not depend on the pitch. Secondly, Fugisaki and Kawashima (65) reported that the relative position of the first two formants were useful in recognizing sounds because every speaker's vocal tract configuration was similar for each sound made. Thus, when the formant frequencies of an utterance were normalized by a shift in frequency, the utterances should have been more similar than before the shift. The sample correlation coefficient (r_{XY}) was calculated to measure the relative improvement in similarity due to the frequency shift described previously.

CHAPTER FOUR

RESULTS

The results of this experiment, in which a determination was made of the usefulness of shifting a voice in frequency in order to improve its machine recognition, were not definitive. This was due to several shortcomings and other limitations inherent in the experiment. Several interesting results were obtained, however, and they will be reported briefly.

A tabulation of the peak sample correlation coefficients chosen revealed that 46.9 per cent of the 4,940 peaks occurred when one of the two voices was shifted. Table 4.1 shows the distribution of sample correlation coefficient peaks for each of the eight shifts and the unshifted position. It should be noted that a shift of one filter or less accounted for 75.4 per cent of all the peak correlation coefficients. A random sample of the peak sample correlation coefficients was averaged together and resulted in an average of 0.733. This average was shown to be statistically greater than a similar value of 0.524. Therefore, in five experiments out of 100, the sample correlation coefficient could be 0.109 larger than 0.524, while the true correlation coefficients were equal. The actual statistical tests are detailed in Appendix B.

A random pair of speakers was chosen to further amplify the results. For speakers one and thirteen, the peak r_{XY} between two

TABLE 4.1 The distribution of peak sample correlation coefficients for all speaker pairs and all letters of the alphabet.

Shift category	Number of peaks
Minus four - first voice shifted down four filters	230
Minus three - first voice shifted down three filters	258
Minus two - first voice shifted down two filters	182
Minus one - first voice shifted down one filter	514
Zero - neither voice shifted	2623
Plus one - first voice shifted up one filter	589
Plus two - first voice shifted up two filters	324
Plus three - first voice shifted three filters up	201
Plus four - first voice shifted up four filters	19

speakers' voices was found in categories zero, plus one, plus two, and plus three. Table 4.2 tabulates the sample correlation coefficients calculated between speakers one and thirteen. The values in each column correspond to the linear relationship between the two speakers' utterances of each letter. The peak correlation coefficient is underlined for each letter. Thirteen peak correlation coefficients appeared in the zero shift category, six in the first shift category, five in the second, and two in the third category. The advantage of employing a uniform shift was then studied. The sample correlation coefficients calculated for each shift for all 26 letters, spoken by speaker one and speaker thirteen, were averaged together. A shift of minus one, zero, plus one, and plus two resulted in average sample correlation coefficients of 0.319, 0.685, 0.646, and 0.497, respectively. A t-test was performed for the two highest average sample correlation coefficients showing that the two means were not statistically different in value. The population of sample correlation coefficients in each category was assumed to be nearly normal, and an F-test supported the assumption that the variances were equal.

A similar study was undertaken to compare the unshifted sample correlation coefficients with the peak sample correlation coefficients for the letter "d". This letter contained approximately an average distribution of the peak sample correlation coefficients among the nine shift categories. The average of the peak sample correlation coefficients was 0.806 and the average of the unshifted sample correlation coefficients was 0.623. Thus, the difference between the unshifted

TABLE 4.2 Sample correlation coefficients in nine shift categories for speaker one and speaker thirteen.

Letter spoken	Shift Category -4	-3	-2	-1	0	+1	+2	+3	+4
a	-.42	-.32	-.04	.52	.92	.76	.37	.01	-.29
b	-.29	-.24	-.14	.10	.62	.73	.70	.26	-.14
c	-.28	-.20	-.06	.20	.54	.79	.86	.31	-.15
d	-.27	-.20	-.04	.29	.69	.73	.67	.20	-.25
e	-.42	-.31	-.13	.09	.49	.64	.51	-.02	-.23
f	-.27	-.21	-.01	.50	.84	.56	.21	-.04	-.16
g	-.30	-.22	.08	.54	.78	.74	.35	-.14	-.29
h	-.33	-.22	-.01	.44	.93	.76	.31	-.02	-.26
i	-.15	.18	.57	.79	.95	.77	.55	.24	-.03
j	-.24	-.20	-.09	.29	.91	.78	.27	.00	-.18
k	-.13	.14	.27	.65	.89	.48	.09	-.15	-.31
l	-.28	-.26	-.18	.08	.54	.83	.72	+.40	.06
m	-.51	-.38	-.25	-.02	.35	.34	.11	-.17	-.42
n	-.35	-.16	+.12	.47	.77	.53	.39	.25	-.20
o	-.19	-.14	+.22	.66	.98	.76	.44	.25	-.03
p	-.29	-.07	.13	.24	.61	.54	.50	.29	-.10
q	-.20	-.08	.07	.31	.79	.82	.60	.23	-.15
r	-.26	-.08	.15	.34	.55	.60	.77	.74	.33
s	-.27	-.27	-.16	.21	.75	.85	.44	+.03	-.19
t	-.18	.08	-.42	.58	.86	.19	.10	-.05	-.20
u	-.20	-.13	.00	.22	.48	.52	.86	.51	-.09
v	-.28	-.25	-.17	.03	.32	.63	.89	.50	-.11
w	-.12	-.01	.21	.70	.94	.58	.23	.12	-.11
x	-.30	-.30	-.20	-.02	.55	.73	.42	.01	-.16
y	-.33	-.21	-.06	.14	.38	.50	.73	.90	.62
z	-.26	-.22	-.15	.02	.38	.60	.81	.42	-.18
average	-.27	-.16	.02	.32	.69	.65	.50	.20	-.11

and peak values was 0.183. A statistical analysis focused on the 87 peak sample correlation coefficients found in a shifted category. The average of these 87 peaks was 0.697, while the average of the corresponding unshifted sample correlation coefficients was 0.296, a difference of 0.401. A t-test was performed to indicate whether this difference was statistically significant. It was found that this difference was significant at a level where in less than one experiment in a thousand, a difference in the sample means this large would be found while the true means were equal. Again the population was assumed normal, and an F-test was performed to indicate whether the variances could be assumed equal.

The remainder of the results obtained for the 190 possible pairs of speakers and the 26 letters were similar to those reported. A final analysis was performed on the twenty voices investigated. The location of the peak sample correlation coefficients was used to indicate the order of voices from the lowest to the highest. This was done in two ways. First of all, each voice was compared to the first speaker and ordered with respect to him. Secondly, the order was determined for each letter and then averaged together. The location of each voice with respect to the others were displayed in Figure 4.1. Finally, a subjective appraisal of the order of the 20 voices was presented in the same figure. In this case, two listeners analyzed the voices and ordered them accordingly.

1) Order relative to speaker one with distances between the voices

lowest highest

15		17		18		2	20	10	3	1	5		7	11	6	14			12
								16	8	4	9								13
								19											

2) Order relative to all other speakers for each letter, averaged together. The relative distances between each voice are shown.

lowest highest

15			18	17		16		19	20	2	8	1		14	6	5	11		13	12
										3		4			7					
										9										
										10										

3) Order relative to a subjective appraisal of each voice by two listeners. The distances between voices are not shown.

lowest highest

10	7	6	3	9	8	2	5	1	4	13	11	12	19	15	16	17	14	20	18

FIGURE 4.1 Three determinations of the order of test voices are the following: 1) order relative to speaker one, 2) order relative to all other speakers for each letter, averaged together, and 3) the order relative to a subjective appraisal of each voice by two listeners.

CHAPTER FIVE

DISCUSSION

The experiment designed to investigate the frequency shift in the power spectrum between pairs of voices was not entirely successful. The results indicated that for nearly one half of the pairs of voices, though, a shift increased the linear relation between the two speakers' utterances. This increase was shown to be significant for the letter "d", and this result could be extended to the other letters. This increase also showed that as a result of a shift, the representations of the utterances of each pair of speakers were found to be more alike. Thus, a frequency shift would probably be valuable to reduce the difference in pitch between two speakers' voices. Unfortunately, the results also indicated that a fixed shift between two voices would not be particularly helpful in reducing this difference in pitch. From the researcher's own experience, however, it seemed that such a fixed shift should have been useful in reducing that difference.

The average value of the peak sample correlation coefficients was encouraging because it did indicate a fairly strong linear relation between representations of each letter spoken by different speakers. Very little can be concluded from its value, however, because of the similarity in the power spectrum for any two utterances. Therefore, valid conclusions must be based on relative values of the coefficients.

The three determinations of the order of voices from lowest to highest were interesting because of the shifts made between two speakers. These shifts should have corresponded with the relative location of each voice when ordered from lowest to highest by the ear. Unfortunately, very little correlation was evident between the ear's analysis and the other two analyses based on the peak sample correlation coefficients.

Several shortcomings of this experiment, in addition to other considerations, affected the results. Because the frequency shift employed was not linear, the results could only be indirectly applied to conclusions concerning actual power spectrum shifts. Secondly, the lower frequency parts of each utterance were given undue emphasis in the calculation of the sample correlation coefficients because of their high relative values. Pre-emphasis of the speech wave and logarithmic amplitude normalization may have reduced this effect. Finally, the processing by the microphone, tape recorder, bandpass filters, rectifiers, and the integrators contributed their respective errors to the whole process. The values thus obtained for the sequence representing an utterance were only approximate.

Several other considerations affected these results as well. The tendency of several speakers' voices to change pitch with each utterance was noted. This tendency and the background noise could have led to the varied results obtained when comparing two speakers. It was also noted that for several utterances, a higher correlation coefficient was obtained when the first formant of an utterance was

matched with a resonance other than the first formant of the other speaker's utterance. A higher correlation coefficient in this instance would probably not correspond to a closer similarity in the two utterances when considered for recognition purposes. These considerations and shortcomings, however, were valuable in formulating conclusions.

CHAPTER SIX

CONCLUSIONS

6.1 EXPERIMENTAL CONCLUSIONS

The experiment conducted as a part of this research contributed to the conclusions reached by this investigator. First of all, the survey of automatic speech recognition systems indicated the need for recognizing the speech of many speakers of both sexes. The analysis of major problems indicated the nature of this problem and the efforts conceived to reduce it. It was found that major differences in speakers' voices could be attributed to the differences in the pitch of their voices. Very little success had been realized, however, from direct attempts to reduce this difference.

This attempt to reduce the difference in voice pitch among speakers was based on a frequency shift of the speech input. Each utterance was shifted with respect to the same utterance spoken by another speaker. The sample correlation coefficient, a measure of the similarity of the two utterances, was calculated at each shift position and the maximum was recorded. This experiment indicated that almost half of those maximum sample correlation coefficients were recorded at a positive or negative shift position. A statistical comparison of the peak sample correlation coefficients with their unshifted counterparts indicated that the average of the peak sample correlation coefficients was significantly larger than the average of

the unshifted sample correlation coefficients. Therefore, this investigator could conclude that some type of normalizing shift of the speech power spectrum for each speaker would be useful in order to enhance the recognition of their speech by machine.

This experiment failed to indicate what type of normalizing shift of the speech power spectrum would be the most useful. The experimental results did not indicate that a fixed shift for the 26 pairs of letters spoken by a pair of speakers would improve recognition accuracy. Finally, the experience gained from processing the speech data proved to be valuable for two reasons. First of all, this experimenter gained a much deeper insight into the nature of speech. Secondly, the problems encountered when recognizing speech were better understood. In this way, the experimental investigation was able to contribute to the overall research effort.

Two other methods were used to accommodate the difference in voice pitch. Forgie and Forgie (58) reported one measure of pitch used to normalize the pitch difference in male and female voices. Separate results were not reported for male and female speakers, although 88 per cent correct recognition was reported for the group's utterances. Li, et al. (61) used a fixed frequency multiplication of 1.16:1 for female compared to male speakers. This transformation, however, did not achieve equal recognition accuracy for male and female speakers. Therefore, efforts have been made to normalize the pitch differences among speakers. The method investigated by this researcher had several things in common with previous methods.

Emphasis was placed on the future implementation of the normalization scheme rather than the optimum frequency shift indicated by study of the speech power spectrum.

It was somewhat difficult to compare the experimental results obtained with the recognition results reported in the literature. It was evident from this research, however, that a frequency shift of type investigated would enhance recognition results for both men and women. Furthermore, recognition would be enhanced for all speakers because the shift investigated could be applied to any pitch difference. Thus, the voices of the same sex would be normalized with respect to each other according to the difference in pitch.

6.2 LITERATURE REVIEW CONCLUSIONS

Many conclusions could be drawn from the research presented here on automatic speech recognition systems. The many efforts reported in the survey of speech recognition systems were only representative of the true effort being exerted to recognize speech. First of all, it was concluded that the accurate recognition of normal speech by a machine was not possible at this time. It was also possible to conclude, however, that progress was being made and that such recognition would be a reality. Already, discrete limited vocabularies were recognized very accurately by machine for several speakers. Furthermore, more and more investigators were turning their attention to the more difficult continuous speech problems.

Each group of speech recognition systems provided the basis for several conclusions this research reached. A limited recognition system, consisting of a small vocabulary, discrete utterances, one speaker, high accuracy, and hardware, could be built according to the template matching approach. If a system was desired with a more simplified hardware approach, the use of zero crossing information would be considered. If a digital computer was available, even greater recognition tasks could be undertaken. Formant frequencies, distinctive features, and spectral features were processed more easily with a computer. Even when using a computer, however, the more complex a recognition task became, the more complex the recognition procedures became. The use of distinctive features also provided a conceptual base to evaluate system performance according to existing physical

characteristics. A formant tracking system, while being very difficult to construct, had a promising future with continuous speech research efforts. Lastly, it was concluded that spectral features, because of their dependence on the power spectrum of speech, would also continue to be valuable in speech recognition research.

The major problems present when recognizing human speech were in addition to those associated with each recognition approach. Several conclusions were apparent from the investigation and discussion of each of the four major problems. The continuous nature of speech was probably the most obvious and well understood problem faced. Solution of this problem depended primarily on the success of segmentation and identification of each segment. The second problem was one of the most complex, but its solution was probably not as critical as the other three in achieving the goal of accurate machine recognition of human speech. From experience with the human recognition system, however, linguistic information may prove to be the difference between accurate recognition and perfect recognition. Thirdly, the problem of variability in repeated utterances by a speaker would best be solved by avoiding features that change with each repeated utterance. The most emphasis was put on the problem of interspeaker variabilities, however, because these variations were so much greater among speakers. The technique of avoiding variable features was not sufficient to solve this problem. Therefore, in addition to other processing, the speech wave will have to be processed to normalize the troublesome differences among speakers. Eventual solutions of these four problems,

fortunately, are not necessarily needed to solve the total speech recognition problem because of the redundant nature of speech communication.

CHAPTER SEVEN

RECOMMENDATIONS

7.1 APPLICATIONS

The literature survey and the experimental investigation were valuable in considering the task of automatic speech recognition by machine. The survey of existing speech recognition efforts revealed not only the state-of-the-art approaches dealing with this task, but also the problems encountered in connection with these approaches. The survey revealed the feasibility of recognizing a discrete limited vocabulary for a small number of speakers. This recognition could be accomplished with nearly perfect accuracy by either a hardware or computerized system. Finally, almost any of the five approaches discussed previously could be used to achieve this recognition accuracy.

The type of speech recognition system described above could be used effectively in many situations. In addition, several systems of this type are available on the market. This type of system was especially useful when the user was unable to hand-tabulate spoken data. One industrial system (12) used a speech recognition system for inspecting television faceplates. The inspector called out the data while positioning the faceplate for measurement. The system automatically typed out the inspection report and listed measurements that were out-of-tolerance. This particular system not only increased

the speed of measurement, but also improved the accuracy of the inspected data.

A speech recognition system would also be valuable when considering the use of remotely-controlled apparatus. The vocabulary could be chosen and used to command almost any apparatus over a telephone or from just across a room. For instance, with an appropriate vocabulary of letters and symbols, a computer program could be submitted from any telephone without a costly terminal. A final limited application could be effected for the deaf or those whose writing or ambulatory skills are handicapped. Of course, a larger system with the capability of recognizing continuous speech would be even more valuable for these tasks and for others. In particular, one of the most useful would be a speech recorder and transcriber.

The literature survey also revealed a wide range of obstacles to the solution of the problem of automated speech recognition. More significantly, however, were the varied schemes adopted to deal with those obstacles. Depending on the particular recognition task, these schemes could be combined to enhance various aspects of speech recognition. These schemes suggested further refinements possible in reducing the obstacles blocking accurate speech recognition. Finally, the analysis of these problems provided new insights for the researcher attempting to work with these problems. Thus, this research could be directly applied to further research on the various problems associated with automatic speech recognition.

The conclusions reached by this investigator's experiment concerning interspeaker differences could also be applied to a speech recognition system. A fixed frequency shift to normalize a speaker's voice pitch to a reference pitch was not indicated as helpful in reducing this difference. Some type of shift was indicated, however, to reduce this difference in the pitch of speakers' voices. These conclusions could be helpful in the designing of methods to solve or circumvent this problem, one of the most serious obstacles blocking accurate speech recognition. In addition, the technique of an automatic shift in the outputs of the analyzing filters rather than a shift of the input voice could be applied to any filter-based system. Such a shift could enhance that system's recognition accuracy for a wider variety of speakers. Another useful experimental idea concerned the use of the spoken alphabet in a recognition system in order to expand the communicating capability of that system.

7.2 FURTHER RESEARCH

The complexity and scope of the automatic speech recognition problem when combined with the usefulness of its solution led to many directions further research could take. One advantage that speech recognition research had was that there were so many problems which are separate and can each contribute individually to the performance of a recognition system. Therefore, it was not necessary to solve the complete problem in order to see the results of each contribution to it.

The next logical step after recognizing and understanding the past approaches and problems with speech recognition would be to design and construct a working system. Once a system was devised, research on each specific problem could be more easily undertaken. In addition, the actual experience with constructing such a system would be invaluable to the understanding of many speech recognition problems. If such a system was to be used primarily as a research tool, then the accessibility and flexibility of a computer processor as noted by Reddy (33) would be a valuable consideration in the design of such a system.

Several basic research needs were noted during this investigation. Probably the first area mentioned dealt with the ideal features associated with speech. These discriminating features either do not exist or have not been discovered yet. Secondly, further investigation was possible in the process of segmenting continuous speech. This speech process was considered analagous to the familiar analog-to-digital

conversion. Another possible broad area of study involved the realization of linguistic aids to the recognition and understanding of speech. This field was particularly attractive due to the many untapped sources of information present in the speech wave. Lastly, research on the nature and range of all the variabilities among speakers and their repeated utterances was called for as necessary to accurate speech recognition by machine.

The experiment conducted among twenty speakers led to several recommendations for further research on the differences in the speech power spectra for speakers. Not only did the nature and range of these differences need to be studied further, but also methods of achieving a reduction in these differences should be investigated. Several methods should be studied to determine the type of shift of the power spectra which would be optimal. Besides the shift described in the experiment, frequency multiplication, linear shifting, and phase multiplication reported by Bogner (66), should be considered. In addition, methods of implementing this normalization technique automatically must be determined. Thus. further research on automatic speech recognition problems could be attempted at many stages of the speech recognition process in order to achieve different effects on the accuracy of its recognition.

BIBLIOGRAPHY

(1) J. L. Flanagan, Speech Analysis, Synthesis, and Perception. New York: Academic Press, 1965.

(2) N. Lindgren, "Machine Recognition of Human Language, Part I - Automatic Speech Recognition," IEEE Spectrum, vol. 2, pp. 114-136, March 1965.

(3) L.C.W. Pols, "Real Time Recognition of Spoken Words," IEEE Trans. Computers, vol. C-20, pp. 972-978, Sept. 1971.

(4) E.E. David, Jr., "Artificial Auditory Recognition in Telephony," IBM J. Research Development, vol. 2, pp. 294-309, Oct. 1958.

(5) P. Denes and M. V. Mathews, "Spoken Digit Recognition Using Time-Frequency Pattern Matching," J. Acoust. Soc. Amer., vol. 32, pp. 1450-1455, Nov. 1960.

(6) R. F. Purton, "Speech Recognition Using Autocorrelation Analysis," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 235-239, June 1968.

(7) M. W. Cannon, Jr., "A Method of Analysis and Recognition for Voiced Vowels," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 154-158, June 1968.

(8) G. L. Clapper, "Automatic Word Recognition," IEEE Spectrum, vol. 8, pp. 57-69, Aug. 1971.

(9) J. H. Warren, "A Pattern Classification Technique for Speech Recognition," IEEE Trans. Audio Electroacoust., vol. AU-19, pp. 281-285, Dec. 1971.

(10) "Automatic Speech Recognition System," Communication News, Jan. 1974.

(11) "Speech Recognition System Handles Service Calls," Electronics, vol. 44, p. 40, Sept. 13, 1971.

(12) "Voice Data Encoding Gives Hands-Free Data Entry," Communication News, p. 42, July 1973.

(13) D. B. Fry and P, Denes, "Mechanical Speech Recognition," in Communication Theory. London: Butterworths Ltd., 1953, pp. 426-432.

(14) D. B. Fry and P. Denes, "On Presenting the Output of a Mechanical Speech Recognizer, "J. Acoust. Soc. Amer., vol. 29, pp. 364-367, March 1957.

(15) H. F. Olson and H. Belar, "Phonetic Typewriter," J. Acoust. Soc. Amer., vol. 28, pp. 1072-1081, Nov. 1956.

(16) H. Dudley and S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," J. Acoust. Soc. Amer., vol. 30, pp. 721-732, Aug. 1958.

(17) G. Sebestyen, "Automatic Recognition of Spoken Numerals," J. Acoust. Soc. Amer., vol. 32, p. 1517, Nov. 1960.

(18) S. R. Petrick and H. M. Willett, "Digital Automatic Word Recognition Procedure," J. Acoust. Soc. Amer., vol. 32, pp. 1516-1517, Nov. 1960.

(19) J. N. Shearme and P. F. Leach, "Some Experiments with a Simple Word Recognition System," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 256-261, June 1968.

(20) M. T. Clark, "Word Recognition by Means of Orthogonal Functions," IEEE Trans. Audio Electroacoust., vol. AU-18, pp. 304-312, Sept. 1970.

(21) A. Ichikawa, Y. Nakano, and K. Nakata, "Evaluation of Various Parameter Sets in Spoken Digits Recognition," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 202-209, June 1973.

(22) L. J. Raphael, "Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word Final Consonants in American English," J. Acoust. Soc. Amer., vol. 51, pp. 1296-1308, April 1972.

(23) E. E. David, Jr. and O. G. Selfridge, "Eyes and Ears for Computers," Proc. IRE, vol. 50, pp. 1093-1101, May 1962.

(24) J. E. Keith Smith and Laura Klem "Vowel Recognition Using a Multiple Discriminant Function," J. Acoust. Soc. Amer., vol. 33, p. 358, March 1961.

(25) L. Gilli and A. R. Meo, "Sequential System for Recognizing Spoken Digits in Real Time," Acustica, vol. 19, pp. 38-48, 1967/1968.

(26) D. J. Comer, "The Use of Waveform Asymmetry to Identify Voiced Sounds," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 500-506, Dec. 1968.

(27) V. I. Velichko and N. G. Zagoruiko, "Automatic Recognition of a Limited Set of Vocal Commands, Abstract 1449," IEEE Audio Electroacoust. Newsletter, p. A-6, Nov. 1972.

(28) J. E. Paul, Jr., "A Limited-Vocabulary, Multi-Speaker Automatic Isolated Word Recognition System, Abstract 1299," IEEE Audio Electroacoust. Newsletter, p. A-5, Nov. 1972.

(29) J. C. Miller, P. W. Ross, and C. M. Wine, "An Adaptive Speech Recognition System Operating in a Remote Time-Shared Computer Environment," IEEE Trans. Audio Electroacoust., vol. AU-18, pp. 26-32, March 1970.

(30) M. Halle and K. Stevens, "Speech Recognition: A Model and a Program for Research," IRE Trans. Information Theory, vol. IT-8, pp. 155-159, Feb. 1962.

(31) R. Alter, "Utilization of Contextual Constraints in Automatic Speech Recognition," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 6-11, March 1968.

(32) A. L. Nelson, M. B. Herscher, T. B. Martin, H. J. Zadell, and J. W. Falter, "Acoustic Recognition by Analog Feature-Abstraction Techniques," Models for the Perception of Speech and Visual Form, Proc. of a Symposium, Cambridge, Mass.: MIT Press, pp. 428-440, 1967.

(33) D. R. Reddy, "Computer Recognition of Connected Speech," J. Acoust. Soc. Amer., vol. 42, pp. 329-347, April 1967.

(34) S. H. Lavington, "Computer Simulation of a Speech Recognition System," Proc. of IEE, vol. 116, pp. 1053-1059, June 1969.

(35) D. R. Reddy, L. D. Erman, and R. B. Neely, "A Model for Machine Recognition of Speech," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 229-238, June 1973.

(36) G. D. Nelson and D. M. Levy, "Selection of Pattern Features by Mathematical Programming Algorithms," IEEE Trans. Systems Science and Cybernetics, vol. SSC-6, pp. 20-25, Jan. 1970.

(37) N. Lindgren, "Machine Recognition of Human Language, Part II - Theoretical Models of Speech Perception and Language," IEEE Spectrum, vol. 2, pp. 44-59, April 1965.

(38) J. Wiren and H. L. Stubbs, "Electronic Binary Selection System for Phoneme Classification," J. Acoust. Soc. Amer., vol. 28, pp. 1082-1091, Nov. 1956.

(39) J. F. Hemdal and G. W. Hughes, "A Feature Based Computer Recognition Program for the Modeling of Vowel Perception," Models for the Perception of Speech and Visual Form, Proc. of a Symposium, Cambridge, Mass.: MIT Press, pp. 440-453, 1967.

(40) S. Itahashi, S. Makino, and K. Kido, "Discrete-Word Recognition Utilizing a Word Dictionary and Phonological Rules," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 239-249, June 1973.

(41) D. G. Bobrow and D. H. Klatt, "A Limited Speech Recognition System," Proc. of AFIPS Fall Joint Computer Conf., vol. 33, pp. 305-318, 1968.

(42) W. A. Lea, "An Approach to Syntactic Recognition Without Phonemics," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 249-258, June 1973.

(43) J. C. R. Licklider and I. P. Pollack, "Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech," J. Acoust. Soc. Amer., vol. 20, pp. 42-51, Jan. 1948.

(44) G. D. Ewing and J. F. Taylor, "Computer Recognition of Speech Using Zero-Crossing Information," IEEE Trans. Audio Electroacoust., vol. AU-17, pp. 37-40, March 1969.

(45) M. R. Ito, "Relationship Between Zero-Crossing Measurements for Speech Analysis and Recognition," J. Acoust. Soc. Amer., vol. 51, pp. 2061-2062, June 1972.

(46) M. R. Ito and R. W. Donaldson, "Zero-Crossing Measurements for Analysis and Recognition of Speech Sounds," IEEE Trans. Audio Electroacoust., vol. AU-19, pp. 235-242, Sept. 1971.

(47) W. Bezdel and H. J. Chandler, "Results of an Analysis and Recognition of Vowels by Computer Using Zero-Crossing Data," Proc. of IEE, vol. 112, pp. 2060-2066, Nov. 1965.

(48) W. Bezdel and J. S. Bridle, "Speech Recognition Using Zero-Crossings Measurements and Sequence Information," Proc. of IEE, vol. 116, pp. 617-623, April 1969.

(49) R. DeMori, "A Descriptive Technique for Automatic Speech Recognition," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 89-100, April 1973.

(50) R. W. A. Scarr, "Zero Crossings as a Means of Obtaining Spectral Information in Speech Analysis," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 247-255, June 1968.

(51) V. Biancomano, "Limited Speech Recognition," QST, vol. 56, pp. 36-39, Oct. 1972.

(52) G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Amer., vol. 24, pp. 175-184, March 1952.

(53) C. F. Teacher, H. G. Kellet and L. R. Focht, "Experimental Limited Vocabulary, Speech Recognizer," IEEE Trans. Audio Electroacoust., vol. AU-15, pp. 127-130, Sept. 1967.

(54) D. I. Durst and M. Lefkowitz, "A Limited-Vocabulary Speech Recognition System," IEEE Student J., vol. 8, pp. 21-29, Sept. 1970.

(55) K. H. Davis, R. Biddulph and S. Balashek, "Automatic Recognition of Spoken Digits," J. Acoust. Soc. Amer., vol. 24, pp. 637-642, Nov. 1952.

(56) R. C. Ohlendorf and C. L. Coates, "Recognition of Spoken Digits Utilizing Sequential Patterns, Abstract 540," IEEE Trans. Audio Electroacoust., vol. AU-17, p. 311, Dec. 1969.

(57) T. G. Von Keller, "An On-Line Recognition System for Spoken Digits," J. Acoust. Soc. Amer., vol. 49, pp. 1288-1296, April 1971.

(58) J. W. Forgie and C. D. Forgie, "Results Obtained from a Vowel Recognition Computer Program," J. Acoust. Soc. Amer., vol. 31, pp. 1480-1489, Nov. 1959.

(59) L. J. Gerstman, "Classification of Self-Normalized Vowels," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 78-80, March 1968.

(60) G. E. Peterson, "Automatic Speech Recognition Procedures," Language and Speech, vol. 4, pp. 200-219, Oct.-Dec. 1961.

(61) K. P. Li, G. W. Hughes and T. B. Snow, "Segment Classification in Continuous Speech," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 50-57, Feb. 1973.

(62) R. W. Schafer and L. R. Rabiner, "Design of Digital Filter Banks for Speech Analysis," Bell System Tech. J., vol. 50, pp. 3097-3115, Dec. 1971.

(63) B. S. Atal and Suzanne L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer., vol. 50, pp. 637-655, Aug. 1971.

(64) P. Denes, "The Design and Operation of the Mechanical Speech Recognizer at University College, London," J. Brit. IRE, vol. 19, pp. 219-234, April 1959.

(65) H. Fujisaki and T. Kawashima, "The Roles of Pitch and Higher Formants in the Perception of Vowels," IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 73-77, March 1968.

(66) R. E. Bogner and J. L. Flanagan, "Frequency Multiplication of Speech Signals," IEEE Trans. Audio Electroacoust., vol. AU-17, pp. 202-208, Sept. 1969.

(67) R. W. Schafer, "A Survey of Digital Speech Processing Techniques," IEEE Trans. Audio Electroacoust., vol. AU-20, pp. 28-35, March 1972.

APPENDIX A. DEFINITION OF TERMS

The following terms are defined or amplified according to the manner in which they were used in this research.

1. ALLOPHONE. An allophone is one of a set of possible variants in the sound of each phoneme.

2. AUTOCORRELATION FUNCTION. The autocorrelation function is defined as the inverse Fourier transform of the power spectrum of the speech sample under consideration.

3. CEPSTRUM. The cepstrum of a speech waveform is the inverse Fourier transform of the logarithm of the speech power spectrum. Ichikawa (21) gave the following equation for the cepstrum C(t):

$$C(t) = \int_{-\infty}^{\infty} \log(P(w)) \cdot \exp(jwt) \cdot dw \qquad t \geq T$$

$$= 0 \qquad t < T$$

where P(w) equals the speech power spectrum and T is the period of speech. Schafer (67) pointed out its advantage, that the vocal tract and excitation components of speech are additive instead of multiplicative. He said that it was especially useful in estimating the pitch period because of this separation.

4. CONDITIONAL PROBABILITY. A conditional probability indicates the probability that one phoneme will follow another in speech. When P(A/B) is that the conditional probability of A relative to B, P(A∩B) is the probability that A and B will occur, and P(B) is the probability that B will occur, then P(A/B) is defined as P(A∩B)/P(B).

5. CONTEXT INFORMATION. Context information in speech is the knowledge imparted to the listener about a speech segment from the sounds surrounding that speech segment considered.
6. DIGRAM FREQUENCY. A digram frequency is the conditional probability that a phoneme will follow another phoneme in speech.
7. DIPTHONGS. A dipthong is a speech sound that starts as one vowel sound and changes to another vowel sound. One example of a dipthong is .
8. EUCLIDEAN DISTANCE. The Euclidean distance between two points in a space is defined as the square root of the sum of the squared distances for each dimension. The Euclidean distance between 011010 and 101110 in six-space is the square root of three.
9. FREQUENCY-AMPLITUDE-TIME TEMPLATE. A frequency-amplitude-time template consists of an M by N matrix of values where M rows correspond to frequency bands and N columns correspond to time periods. Each value of the matrix reflects the amplitude of the frequency content in the frequency band and time period associated with that position in the matrix.
10. FORMANT. A formant is one of several vocal tract resonances which appear in the power spectrum of speech.
11. GLOTTAL EXCITATION. The glottal excitation is the sound stimulum produced at the vocal cords.

12. HAMMING DISTANCE. The Hamming distance is a measure of the distance between two points in a space. For instance, in six-space the Hamming distance between 011010 and 101110 is three because there are three binary places that are different.

13. INFORMATION. Information is the knowledge or facts communicated about the speech under consideration. For instance, spectral information is that knowledge imparted to the listener by the frequencies present in the speech wave, while zero crossing information is that knowledge provided by the zero crossings of the speech wave.

14. INTERMEDIATE PHONEME RECOGNITION. Intermediate phoneme recognition is a method of recognizing speech by first recognizing the individual phonemes making up speech. The acutal words of speech are recognized from a string of phonemes.

15. LINEAR PREDICITVE COEFFICIENTS. A set of linear predictive coefficients are calculated on the basis of past speech samples in order to predict another speech sample. This calculation is based on the least square error concept and is done by solving simultaneous linear equations.

16. LINGUISTICS. Linguistics is the study of a language of human speech. In this research linguisitcs is the study of information contained in speech other than the acoustic information. The linguistic information is related to the language used and both the listener and the speaker depend on it for human speech recognition.

17. LOW-ORDER SPECTRAL MOMENT ANALYSIS. Low order spectral moment analysis is a formant estimation technique based on the mean, variance, and higher-order moments of the speech wave.

18. NORMALIZATION. Normalization is a process in which measures of speech are transformed in order to make them more regular. Amplitude normalization consists of referring every utterance to a certain fixed level of average or peak intensity. Time normalization stretches or shrinks an utterance in time so that every utterance has the same length. Frequency normalization of a voice provides a shift of each voice pitch to a standard frequency or may require that the distance between two formants be the same.

19. OCTAVE FILTER. A set of octave filters is a set of bandpass filters whose ratio of bandwidth to center frequency for each filter remains constant.

20. OPPOSITION. An opposition describes a speech condition where one of two contrary conditions exist. Lindgren (37) said that there are 12 oppositions for all human speech. The 12 oppositions used in all speech are given in Table A.1 while the pattern that English phonemes make in nine of these oppositions is given in Table A.2.

21. ORTHOGONAL BASIS. An orthogonal basis for the representation of speech is one in which the equations representing speech are statistically independent or mutually perpendicular.

TABLE A.1 Characteristics of the twelve oppositions of human speech. Used by permission (37).

Opposition	Characteristics
vocalic vs. nonvocalic	Presence versus absence of a sharply defined formant structure.
consonantal vs. nonconsonantal	Low versus high total energy.
compact vs. diffuse	Higher versus lower concentration of energy (intensity) in a relatively narrow, central region of the spectrum, accompanied by an increase (vs. decrease) of the total energy.
tense vs. lax	Higher versus lower total energy in conjunction with a greater versus smaller spread of the spectrum and in time.
voiced vs. voiceless	Presence versus absence of periodic low-frequency excitation.
nasal vs. oral	Spreading the available energy over wider (vs. narrower) frequency regions by a reduction in the intensity of certain (primarily the first) formants and introduction of additional (nasal) formants.
discontinuous vs. continuant	Silence followed and/or preceded by spread of energy over a wide frequency region (either as a burst or a rapid transition of vowel formants) vs. absence of abrupt transition between sound and such silence.
strident vs. mellow	Higher intensity noise vs. low intensity noise.
checked vs. unchecked	Higher rate of discharge of energy within a reduced interval of time versus lower rate of discharge within a longer interval.
grave vs. acute	Concentration of energy in the lower (vs. upper) frequencies of the spectrum.
flat vs. plain	Flat phonemes in contradistinction to the corresponding plain ones are characterized by a downward shift or weakening of some of their upper-frequency components.
sharp vs. plain	Sharp phonemes in contradistinction to the corresponding plain ones are characterized by an upward shift of some of their frequency components.

TABLE A.2 English phonemes in relation to nine universal oppositions. Used by permission (37).

Opposition	Phonemes																						
	ɑ	æ	ɛ	ʊ	ʌ	ɪ	l	ŋ	ʃ	tʃ	k	ʒ	dʒ	g	m	f	p	v	b	n	s	θ	t	z	ð	d	h
+vocalic -nonvocalic	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+consonantal -nonconsonantal	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
+compact -diffuse	+	+	+	-	-	-		+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	
+tense -lax									+	+	+	-	-	-		+	+	-	-		+	+	+	-	-	-	+
+nasal -oral								+	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	
+discontinuous -continuant									-	+	+	-	+	+		-	+	-	+		-	-	+	-	-	+	
+strident -mellow									+	+	-	+	+	-							+	-	-	+	-	-	
+grave -acute	+	+	-	+	+	-									+	+	+	+	+	-	-	-	-	-	-	-	
+flat -plain	+	-		+	-																						

22. PARTIAL AUTOCORRELATION COEFFICIENTS. A partial autocorrelation coefficient is defined by Ichikawa (21) as the correlation coefficient calculated between two speech samples after the influence due to the samples between the two is removed.

23. PHONEME. A phoneme is the smallest unit of speech and corresponds to individual sounds. A list of the English phonemes, their symbols, and a representative sound for each phoneme are given in Table A.3.

24. PHONOLOGICAL CONTEXT. Phonological context is the sound environment on either side of a speech sound.

25. PITCH. Flanagan (1) said, "Pitch is that subjective attribute which admits of a rank ordering on a scale ranging from low to high." Another more stringent definition is that the pitch of a voice is directly related to the frequency of glottal excitation. In the context of this research, the pitch of a voice was considered to correspond closely with this frequency although it was also affected by the entire speech sound.

26. PITCH SYNCHRONOUS ANALYSIS. Pitch synchronous analysis was defined by Reddy (33) as expanding the pitch period of a speech sample by the Fourier series expansion. This analysis was used to determine pitch and the envelope of the speech power spectrum.

27. POWER SPECTRUM. The power spectrum or power spectral density for a speech sound is the distribution of average power as a function of frequency for all frequencies present in the speech.

TABLE A.3 Phonemes of the English language. Used by permission (2).

Phonetic symbol	Key word	Phonetic symbol	Key word
Simple vowels		**Plosives**	
I	fit	b	bad
i	feet	d	dive
ɛ	let	g	give
æ	bat	p	pot
ʌ	but	t	toy
ɑ	not	k	cat
ɔ	law	**Nasal consonants**	
U	book	m	may
u	boot	n	now
	bird	ŋ	sing
	Bert	**Affricatives**	
Complex vowels		tʃ	church
e	pain	dʒ	judge
o	go	**Fricatives**	
aU	house	z	zero
ɔI	boy	ʒ	vision
IU	few	v	very
Semivowels and liquids		ð	that
j	you	h	hat
w	we	f	fat
l	late	θ	thing
r	rate	ʃ	shed
		s	sat

28. PROSODEME. A prosodeme is a measure of speech related to duration, rhythm, stress, intonation, and accent. Prosodic parameters are measures of these attributes, the prosody of a speech utterance.

29. RESONANCE. A resonance is a large vibration due to a small excitation of nearly the same frequency or multiple thereof. In speech, the vocal tract produces the natural conditions contributing to resonances called formants.

30. SEGMENTATION. Segmentation is a process carried out on the continuous speech wave. This process divides speech into the smallest speech units being recognized, usually phonemes or words.

31. SEMANTICS. Semantics is the study of the meanings associated with words and phrases and their context. Semantic information is part of the linguistic information which is used in human speech recognition.

32. SPECTRAL FEATURE. A spectral feature of speech is any parameter extracted from or corresponding to the speech power system.

33. SYNTAX. Syntax is the section of grammar which deals with the methods of connection of words into clauses, phrases, sentences, and paragraphs.

34. TEMPLATE. A template is a fixed pattern or overlay with which the input speech is compared. The form of the template is determined by the utterances to be used as a comparison to the input utterance.

35. WAVEFORM ASYMMETRY. Waveform asymmetry is a characteristic of the voiced speech waveform when the peaks above or below the

axis are greater than the opposite peaks. The area under the two curves on each side of the axis remains the same, however.

B. Statistical Tests

1. SIGNIFICANCE OF THE VALUE OF THE AVERAGE CORRELATION COEFFICIENT

The Fisher r to Z transformation was used to calculate the value of the true correlation coefficient (ρ_o) which was significantly less than the average sample correlation coefficient (r_{XY}). The distribution of the points correlated was assumed to be nearly bivariate normal; this was a good assumption in this situation because of the size and nature of the sample. For a value of r_{XY} equal to 0.733, Z would equal 0.935. The test statistic was equal to $(Z - \delta)(N - 3)^{\frac{1}{2}}$ where N was the number of samples correlated and δ was the transformed ρ_o. This test statistic was compared with a normal distribution and had to be greater than 1.645 at the .05 level in order to reject the hypothesis that the value of r_{XY} was statistically the same as ρ_o. Solving the equation, $(Z - \delta)(N - 3)^{\frac{1}{2}}$ greater than or equal to 1.645, for δ resulting in a δ equal to 0.582, corresponding to a ρ_o equal to 0.524. Therefore, an r_{XY} equal to 0.733 was statistically greater than a ρ_o equal to 0.524 according to the conditions of the test.

2. COMPARISON OF THE SAMPLE CORRELATION COEFFICIENTS BETWEEN SPEAKERS ONE AND THIRTEEN AT TWO DIFFERENT SHIFTS.

An F-test was performed on the sample variances around the mean of sample correlation coefficients for no shift and for a shift of one. The distribution of the sample correlation coefficients was assumed to be normal at this distance from 1.000. At the .05 level of significance, the F-statistic value of 1.58 was less than the

F-value at that level, 1.95. Therefore, this test failed to reject the hypothesis that the two true variances were equal. Thus, the assumption that the two variances were equal was valid.

A t-test was performed to test the hypothesis that the two means of sample correlation coefficients were the same. The two shift categories evaluated with this test are categories zero and one in Table 4.2. The value of the calculated t-statistic was 0.729 which corresponded to a .20 level of significance. Therefore, the test failed to reject the hypothesis that the two means were the same at the .05 level of significance.

3. COMPARISON OF THE PEAK SAMPLE CORRELATION COEFFICIENTS AND THE UNSHIFTED SAMPLE CORRELATION COEFFICIENTS FOR THE LETTER "D".

An F-test was performed on the sample variances calculated from the unshifted and shifted sample correlation coefficients. The population of sample correlation coefficients was assumed to be normally distributed. The F-value calculated, 1.03, had to be greater than the corresponding F-value at the .05 level, 1.45, in order to reject the hypothesis that the two true variances were equal. Thus, the assumption that the two variances were equal was valid.

A t-test was performed in the same manner as in part 2 of B. A t-value of 9.56 was calculated and this value was greater than the standard t-value at the .001 level, 3.09. Thus, the hypothesis that there was no difference in the average value of the peak sample correlation coefficients compared to the unshifted sample correlation coefficients was rejected.